

UNITED STATES AIR FORCE  
SUMMER RESEARCH PROGRAM -- 1996  
SUMMER RESEARCH EXTENSION PROGRAM FINAL REPORTS

VOLUME 4B  
WRIGHT LABORATORY

RESEARCH & DEVELOPMENT LABORATORIES  
5800 Uplander Way  
Culver City, CA 90230-6608

Program Director, RDL  
Gary Moore

Program Manager, AFOSR  
Major Linda Steel-Goodwin

Program Manager, RDL  
Scott Licoscas

Program Administrator, RDL  
Johnetta Thompson

Program Administrator  
Rebecca Kelly-Clemmons

Submitted to:

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH  
Bolling Air Force Base  
Washington, D.C.  
December 1996

20010319 012

AQM01-06-1067

# REPORT DOCUMENTATION PAGE

AFRL-SR-BL-TR-00-

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering the data, reviewing and collecting the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Project, Washington, DC 20503.

Reviewing  
Information

0706

1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE December, 1996		3. REPORT TYPE AND DATES COVERED	
4. TITLE AND SUBTITLE 1996 Summer Research Program (SRP), Summer Research Extension Program (SREP), Final Report, Volume 4B, Wright Laboratory				5. FUNDING NUMBERS F49620-93-C-0063	
6. AUTHOR(S) Gary Moore					
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Research & Development Laboratories (RDL) 5800 Uplander Way Culver City, CA 90230-6608				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Office of Scientific Research (AFOSR) 801 N. Randolph St. Arlington, VA 22203-1977				10. SPONSORING/MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES					
12a. DISTRIBUTION AVAILABILITY STATEMENT Approved for Public Release				12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words) The United States Air Force Summer Research Program (SRP) is designed to introduce university, college, and technical institute faculty members to Air Force research. This is accomplished by the faculty members, graduate students, and high school students being selected on a nationally advertised competitive basis during the summer intersession period to perform research at Air Force Research Laboratory (AFRL) Technical Directorates and Air Force Air Logistics Centers (ALC). AFOSR also offers its research associates (faculty only) an opportunity, under the Summer Research Extension Program (SREP), to continue their AFOSR-sponsored research at their home institutions through the award of research grants. This volume consists of a listing of the participants for the SREP and the technical report from each participant working at the AFRL Wright Laboratory.					
14. SUBJECT TERMS Air Force Research, Air Force, Engineering, Laboratories, Reports, Summer, Universities, Faculty, Graduate Student, High School Student				15. NUMBER OF PAGES	
				16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UL		

## GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to **stay within the lines** to meet **optical scanning requirements**.

**Block 1. Agency Use Only** (*Leave blank*).

**Block 2. Report Date.** Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year.

**Block 3. Type of Report and Dates Covered.** State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

**Block 4. Title and Subtitle.** A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

**Block 5. Funding Numbers.** To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

**C** - Contract  
**G** - Grant  
**PE** - Program  
Element

**PR** - Project  
**TA** - Task  
**WU** - Work Unit  
Accession No.

**Block 6. Author(s).** Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

**Block 7. Performing Organization Name(s) and Address(es).** Self-explanatory.

**Block 8. Performing Organization Report Number.** Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

**Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es).** Self-explanatory.

**Block 10. Sponsoring/Monitoring Agency Report Number.** (*If known*)

**Block 11. Supplementary Notes.** Enter information not included elsewhere such as: Prepared in cooperation with....; Trans. of....; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

**Block 12a. Distribution/Availability Statement.** Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NDFDRN, REL, ITAR).

**DOD** - See DoDD 5230.24, "Distribution Statements on Technical Documents."

**DOE** - See authorities.

**NASA** - See Handbook NHB 2200.2.

**NTIS** - Leave blank.

**Block 12b. Distribution Code.**

**000** - Leave blank.

**DOE** - Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.

Leave blank.

**NASA** - Leave blank.

**NTIS** -

**Block 13. Abstract.** Include a brief (*Maximum 200 words*) factual summary of the most significant information contained in the report.

**Block 14. Subject Terms.** Keywords or phrases identifying major subjects in the report.

**Block 15. Number of Pages.** Enter the total number of pages.

**Block 16. Price Code.** Enter appropriate price code (*NTIS only*).

**Blocks 17. - 19. Security Classifications.** Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

**Block 20. Limitation of Abstract.** This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.

## **PREFACE**

This volume is part of a five-volume set that summarizes the research of participants in the 1996 AFOSR Summer Research Extension Program (SREP.) The current volume, Volume 1 of 5, presents the final reports of SREP participants at Armstrong Laboratory. Volume 1 also includes the Management Report.

Reports presented in this volume are arranged alphabetically by author and are numbered consecutively – e.g., 1-1, 1-2, 1-3; 2-1, 2-2, 2-3, with each series of reports preceded by a 35 page management summary. Reports in the five-volume set are organized as follows:

<b>VOLUME</b>	<b>TITLE</b>
1	Armstrong Laboratory
2	Phillips Laboratory
3	Rome Laboratory
4A	Wright Laboratory
4B	Wright Laboratory
5	Arnold Engineering Development Center Air Logistics Centers



# 1996 SREP FINAL REPORTS

## Armstrong Laboratory

### VOLUME 1

<b>Report #</b>	<b>Report Title Author's University</b>	<b>Report Author</b>
1	<b>Chlorinated Ethene Transformation, Sorption &amp; Product Distr in Metallic Iron/Water Systems: Effect of Iron Properties</b> Washington State University, Pullman, WA	<b>Dr. Richelle M Allen-King</b> Dept. of Geology AL/EQ
2	<b>Dynamically Adaptive Interfaces: A Preliminary Investigation</b> Wright State University, Dayton, OH	<b>Dr. Kevin B Bennett</b> Dept. of Psychology AL/CF
3	<b>Geographically Distributed Collaborative Work Environment</b> California State University, Hayward, CA	<b>Dr. Alexander B Bordetsky</b> Dept. Decesion Sciences AL/HR
4	<b>Development of Fluorescence Post Labeling Assay for DNA Adducts: Chloroacetaldeh</b> New York Univ Dental/Medical School, New York, NY	<b>Dr. Joseph B Guttenplan</b> Dept. of Chemistry AL/OE
5	<b>The Checkmark Pattern &amp; Regression to the Mean in Dioxin Half Life Studies</b> University of South Alabama, Mobile, AL	<b>Dr. Pandurang M Kulkarni</b> Dept. of Statistics AL/AO
6	<b>Determination of the Enzymatic Constraints Limiting the Growth of Pseudomonas</b> University of Dayton, Dayton, OH	<b>Dr. Michael P Labare</b> Dept. of Marine Sciences AL/HR
7	<b>Tuned Selectivity Solid Phase Microextraction</b> Clarkson University, Potsdam, NY	<b>Dr. Barry K Lavine</b> Dept. of Chemistry AL/EQ
8	<b>A Cognitive Engineering Approach to Distributed Team Decision Making During</b> University of Georgia, Athens, GA	<b>Dr. Robert P Mahan</b> Dept. of Psychology AL/CF
9	<b>Repetative Sequence Based PCR: An Epidemiological Study of a Streptococcus</b> Stonehill College, North Easton, MA	<b>Dr. Sandra McAlister</b> Dept. of Biology AL/CF
10	<b>An Investigation into the Efficacy of Headphone Listening for Localization of</b> Middle Tennessee State University, Murfreesbord, TN	<b>Dr. Alan D. Musicant</b> Dept. of Psychology AL/CF
11	<b>The Neck Models to Predict Human Tolerance in a G-Y</b> CUNY-City College, New York, NY	<b>Dr. Ali M. Sadegh</b> Dept. of Mech Engineering AL/CF

## 1996 SREP FINAL REPORTS

Armstrong Laboratory

### VOLUME 1 (cont.)

<b>Report #</b>	<b>Report Title Author's University</b>	<b>Report Author</b>
12	Tracer Methodology Development for Enhanced Passive Ventilation for Soil University of Florida, Gainesville, FL	Dr. William R. Wise Dept. of Civil Engineering AL/EQ
13	Application of a Distribution-Based Assessment of Mission Readiness System for the Evaluation of Personnel Training Texas A&M University, College Station, TX	Dr. David J. Woehr Dept. of Psychology AL/HR
14	Electrophysiological, Behavioral, and Subjective Indexes of Workload when Performing Multiple Tasks Washington State University, Pullman, WA	Ms. Lisa Fournier Dept. of Psychology AL/CF
15	Methods for Establishing Design Limits to Ensure Accomodation for Ergonomic Design Miami University, Oxford, OH	Ms. Kristie Nemeth Dept. of Psychology AL/HR

# 1996 SREP FINAL REPORTS

## Phillips Laboratory

### VOLUME 2

<b>Report #</b>	<b>Report Title Author's University</b>	<b>Report Author</b>
1	Experimental Study of the Tilt Angular Anisotropy Correlation & the Effect Georgia Tech Research Institute, Atlanta, GA	Dr. Mikhail Belen'kii Dept. of Electro Optics PL/LI
2	Performance Evaluations & Computer Simulations of Synchronous & Asynchronous California State University, Fresno, CA	Dr. Daniel C. Bukofzer Dept. of Elec Engineering PL/VT
3	MM4 Model Experiments on the Effects of Cloud Shading Texas Tech University, Lubbock, TX	Dr. Chia-Bo Chang Dept. of Geosciences PL/GP
4	Miniature Laser Gyro consisting in a Pair of Unidirectional Ring Lasers University of New Mexico, Albuquerque, NM	Dr. Jean-Claude M. Diels Dept. of Physics PL/LI
5	Simulations & Theoretical Studies of Ultrafast Silicon Avalanche Old Dominion University, Norfolk, VA	Dr. Ravindra P. Joshi Dept. of Elec Engineering PL/WS
6	Theory of Wave Propagation in a Time-Varying Magnetoplasma Medium & Applications to Geophysical Phenomena University of Massachusetts Lowell, Lowell, MA	Dr. Dikshitulu K. Kalluri Dept. of Elec Engineering PL/GP
7	Thermal Analysis for the Applications of High Power Lasers in Large-Area Materials Processing University of Central Florida, Orlando, FL	Dr. Arvinda Kar Dept. of Engineering PL/LI
8	Analytical Noise Modeling and Optimization of a Phasor-Based Phase Texas Tech University, Lubbock, TX	Dr. Thomas F. Krile Dept. of Elec Engineering PL/LI
9	Mathematical Modeling of Thermionic-AMTEC Cascade System for Space Power Texas Tech University, Lubbock, TX	Dr. M. Arfin K. Lodhi Dept. of Physics PL/VT
10	Preparation & characterization of Polymer Blends Ohio State University, Columbus, OH	Dr. Charles J. Noel Dept. of Chemistry PL/RK
11	Evaluation of Particle & Energy Transport to Anode, Cathode University of Texas-Denton, Denton, TX	Dr. Carlos A. Ordonez Dept. of Physics PL/WS
12	Analysis of the Structure & Motion of Equatorial Emission Depletion Bands Using Optical All-Sky Images University of Massachusetts Lowell, Lowell, MA	Dr. Ronald M. Pickett Dept. of Psychology PL/GP

# 1996 SREP FINAL REPORTS

Phillips Laboratory

## VOLUME 2 (cont.)

<u>Report #</u>	<u>Author's University</u>	<u>Report Author</u>
13.	<b>On the Fluid Dynamics of High Pressure Atomization in Rocket Propulsion</b> University of Illinois-Chicago, Chicago, IL	<b>Dr. Dimos Poulikakos</b> Dept. of Mech Engineering PL/RK
14	<b>Gigahertz Modulation &amp; Ultrafast Gain Build-up in Iodine Lasers</b> University of New Mexico, Albuquerque, NM	<b>Dr. W. Rudolph</b> Dept. of Physics PL/LI
15	<b>Inversion of Hyperspectral Atmospheric Radiance Images for the Measurement of Temperature, Turbulence, and Velocity</b> University of New Mexico, Albuquerque, NM	<b>Dr. David Watt</b> Dept. of Mech Engineering PL/GP

# 1996 SREP FINAL REPORTS

Rome Laboratory

## VOLUME 3

<b>Report #</b>	<b>Author's University</b>	<b>Report Author</b>
1	Performance Analysis of an ATM-Satellite System Florida Atlantic University, Boca Raton, FL	Dr. Valentine Aalo Dept. of Elec Engineering RL/C3
2	Reformulating Domain Theories to Improve their Computational Usefulness Oklahoma State University, Stillwater, OK	Dr. David P. Benjamin Dept. of Comp Engineering RL/C3
3	An Analysis of the Adaptive Displaced Phase Centered Antenna Lehigh University, Bethlehem, PA	Dr. Rick S. Blum Dept. Elec Engineering RL/OC
4	Effect of Concatenated Codes on the Transport of ATM-Based Traffic California Polytechnic State, San Luis Obispo, CA	Dr. Mostafa Chinichian Dept. of Engineering RL/C3
5	Development of Efficient Algorithms & Software Codes for Lossless and Near-Lossless Compression of Digitized Images Oakland University, Rochester, MI	Dr. Manohar K. Das Dept. Elec Engineering RL/IR
6	Mode-Locked Fiber Lasers Rensselaer Polytechnic Institution, Troy, NY	Dr. Joseph W. Haus Dept. of Physics RL/OC
7	Magnitude & Phase Measurements of Electromagnetic Fields Using Infrared University of Colorado, Colorado Springs, CO	Dr. John D. Norgard Dept. Elec Engineering RL/ER
8	Image Multiresolution Decomposition & Progressive Transmission Using Wavelets New Jersey Institute of Technology, Newark, NJ	Dr. Frank Y. Shih Dept. of Comp Science RL/IR
9	Investigation of Si-Based Quantum Well Intersubband Lasers University of Massachusetts-Boston, Boston, MA	Dr. Gang Sun Dept. of Physics RL/ER
10	Numerical Study of Bistatic Scattering from Land Surfaces at Grazing Incidence Oklahoma State University, Stillwater, OK	Dr. James C. West Dept. of Elec Engineering RL/ER

# 1996 SREP FINAL REPORTS

## Wright Laboratory

### VOLUME 4A

Report #	Author's University	Report Author
1	Barrel-Launched Adaptive Munition Experimental Round Research Auburn University, Auburn, AL	Dr. Ronald M. Barrett Dept. of Aerospace Eng WL/MN
2	Modeling & Design of New Cold Cathode Emitters & Photocathodes University of Cincinnati, Cincinnati, OH	Dr. Marc M. Cahay Dept. of Elec Engineering WL/EL
3	Unsteady Aerodynamics University of California-Berkeley, Berkeley, CA	Dr. Gary Chapman Dept. of Aerospace Eng WL/MN
4	Characteristics of the Texture Formed During the Annealing of Copper Plate University of Nebraska-Lincoln, Lincoln, NE	Dr. Robert J. DeAngelis Dept. of Mech Engineering WL/MN
5	Development of Perturbed Photoreflectance, Implementation of Nonlinear Optical Parametric Devices Bowling Green State University	Dr. Yujie J. Ding Dept. of Physics WL/EL
6	Computations of Drag Reduction & Boundary Layer Structure on a Turbine Blade with an Oscillating Bleed Flow University of Dayton, Dayton, OH	Dr. Elizabeth A. Ervin Dept. of Mech Engineering WL/PO
7	Low Signal to Noise Signal Processor for Laser Doppler Velocimetry North Carolina State University, Raleigh, NC	Dr. Richard D. Gould Dept. of Mech Engineering WL/PO
8	Modeling & Control for Rotating Stall in Aeroengines Louisiana State University, Baton Rouge, LA	Dr. Guoxiang Gu Dept. of Elec Engineering WL/FI
9	Scaleable Parallel Processing for Real-time Rule-Based Decision Aids University of Missouri-Columbia, Columbia, MO	Dr. Chun-Shin Lin Dept. of Elec Engineering WL/FI
10	Quantitative Image Location & Processing in Ballistic Holograms University of West Florida, Pensacola, FL	Dr. James S. Marsh Dept. of Physics WL/MN
11	Experimental & Computational Investigation of Flame Suppression University of North Texas, Denton, TX	Dr. Paul Marshall Dept. of Chemistry WL/ML
12	Investigations of Shear Localization in Energetic Materials Systems University of Notre Dame, Notre Dame, IN	Dr. James J. Mason Dept. of Aerospace Eng WL/MN

# 1996 SREP FINAL REPORTS

Wright Laboratory

VOLUME 4A (cont.)

<b>Report #</b>	<b>Author's University</b>	<b>Report Author</b>
13	<b>A Time Slotted Approach to Real-Time Message Scheduling on SCI University of Nebraska-Lincoln, Lincoln, NE</b>	<b>Dr. Sarit Mukherjee Dept. of Comp Engineering WL/AA</b>
14	<b>Dielectric Resonator Measurements on High Temperature Superconductor (HTS) Wright State University, Dayton, OH</b>	<b>Dr. Krishna Naishadham Dept. Elec Engineering WL/ML</b>
15	<b>Modeling of Initiation &amp; Propagation of Detonation Energetic Solids University of Notre Dame, Notre Dame, IN</b>	<b>Dr. Joseph M. Powers Dept. of Aerospace WL/MN</b>
16	<b>Robust control Design for Nonlinear Uncertain Systems by Merging University of Central Florida, Orlando, FL</b>	<b>Dr. Zhihua Qu Dept. of Elec Engineering WL/MN</b>

# 1996 SREP FINAL REPORTS

Wright Laboratory

## VOLUME 4B

Report #	Author's University	Report Author
17	HELPR: A Hybrid Evolutionary Learning System Wright State University, Dayton, OH	Dr. Mateen M. Rizki Dept. of Comp Engineering WL/AA
18	Virtual Materials Processing: automated Fixture Design for Materials Southern Illinois University-Carbondale, IL	Dr. Yiming K. Rong Dept. of Technology WL/ML
19	A Flexible Architecture for Communication Systems (FACS): Software AM Radio Wright State University, Dayton, OH	Dr. John L. Schmalzel Dept. of Engineering WL/AA
20	A Design Strategy for Preventing High Cycle Fatigue by Minimizing Sensitivity of Bladed Disks to Mistuning Wright State University, Dayton, OH	Dr. Joseph C. Slater Dept. of Mech Engineering WL/FI
21	Growth of Silicon Carbide Thin Films by Molecular Beam Epitaxy University of Cincinnati, Cincinnati, OH	Dr. Andrew J. Steckl Dept. of Elec Engineering WL/FI
22	Performance of Iterative & Noniterative Schemes for Image Restoration University of Arizona, Tucson, AZ	Dr. Malur K. Sundareshan Dept. of Elec Engineering WL/MN
23	Improving the Tribological Properties of Hard TiC Coatings University of New Orleans, New Orleans, LA	Dr. Jinke Tang Dept. of Physics WL/ML
24	Development of Massively Parallel Epic Hydrocode in Cray T3D Using PVM Florida Atlantic University, Boca Raton, FL	Dr. Chi-Tay Tsai Dept. of Mech Engineering WL/MN
25	Supramolecular Multilayer Assemblies w/Periodicities in a Submicron Range Western Michigan University, Kalamazoo, MI	Dr. Vladimir V. Tsukruk Dept. of Physics WL/ML
26	Distributed Control of Nonlinear Flexible Beams & Plates w/Mechanical & Temperature Excitations University of Kentucky, Lexington, KY	Dr. Horn-Sen Tzou Dept. of Mech Engineering WL/FI
27	A Progressive Refinement Approach to Planning & Scheduling University of Colorado-Denver, Denver, CO	Dr. William J. Wolfe Dept. of Comp Engineering WL/MT
28	Development of a New Numerical Boundary condition for Perfect Conductors University of Idaho, Moscow, OH	Dr. Jeffrey L. Young Dept. of Elec Engineering WL/FI



# 1996 SREP FINAL REPORTS

Wright Laboratory

## VOLUME 4B (cont.)

<b>Report #</b>	<b>Author's University</b>	<b>Report Author</b>
29	<b>Eigenstructure Assignment in Missile Autopilot Design Using a Unified Spectral Louisiana State University, Baton Rouge, LA</b>	<b>Dr. Jianchao Zhu Dept. of Elec Engineering WL/FI</b>
30	<b>Design &amp; Implementation of a GNSS Software Radio Receiver Ohio University, Athens, OH</b>	<b>Dr. Dennis M. Akos Dept. of Elec Engineering</b>
31	<b>Experimental &amp; Numerical Study of Localized Shear as an Initiation Mechanism University of Notre Dame, Notre Dame, IN</b>	<b>Mr. Richard J. Caspar Dept. of Aero Engineering WL/MN</b>
32	<b>A Molecular-Level view of Solvation in Supercritical Fluid Systems State University of New York – Buffalo, Buffalo, NY</b>	<b>Ms. Emily D. Niemeyer Dept. of Chemistry WL/PO</b>
33	<b>Initiation of Explosives by High Shear Strain Rate Impact University of Notre Dame, Notre Dame, IN</b>	<b>Mr. Keith M. Roessig Dept. of Aero Engineering WL/MN</b>

# 1996 SREP FINAL REPORTS

## VOLUME 5

<u>Report #</u>	<u>Author's University</u>	<u>Report Author</u>
-----------------	----------------------------	----------------------

### Arnold Engineering Development Center

1	Facility Health Monitoring & Diagnosis Vanderbilt University, Nashville, TN	Dr. Theodore Bapty Dept. of Elec Engineering AEDC
---	--	---

### Air Logistic Centers

2	Fatigue Crack Growth Rates in Naturally-Coroded Aircraft Aluminum University of Oklahome, Norman, OK	Dr. James D. Baldwin Dept. of Mech Engineering OCALC
3	A Novel Artificial Neural Network Classifier for Multi-Modal University of Toledo, Toledo, OH	Dr. Gursel Serpen Dept. of Elec Engineering OOALC
4	Development of a Cost-Effective Organizational Information System West Virginia University, Morgantown, WV	Dr. Michael D. Wolfe Dept. Mgmt Science SAALC
5	Implementation of a Scheduling Software w/Shop Floor Parts Tracking Sys University of Wisconsin-Stout, Menomonic, WI	Dr. Norman D. Zhou Dept. of Technology SMALC
6	Development of a High Performance Electric Vehicle Actuator System Clarkson University, Potsdam, NY	Dr. James J. Carroll Dept. Elec Engineering WRALC

# **Hybrid Evolutionary Learning System**

**Mateen M. Rizki**  
Associate Professor  
Department of Computer Science and Engineering

College of Engineering and Computer Science  
Wright State University  
Dayton, Ohio 45435

Final Report for:  
Summer Research Extension Program  
Avionics Laboratory (WL/AA)

Sponsored by:  
Air Force Office of Scientific Research  
Bolling Air Force Base, DC

and

Avionics Laboratory  
Wright-Patterson Air Force Base, Dayton Ohio

December 1996

# Hybrid Evolutionary Learning System

Mateen M. Rizki  
Associate Professor  
Department of Computer Science and Engineering  
Wright State University

## Abstract

E-MORPH is a multi-phase evolutionary learning system that evolves cooperative sets of feature detectors and combines their response using a simple nearest neighbor classifier to form a complete pattern recognition system. The learning system evolves registered sets of primitive morphological detectors that directly measure normalized radar signatures. Special convolution kernels are evolved to extract information from the output of the primitive transforms to form real valued feature vectors. Starting with a population of trivial randomly generated transforms, EMORPH uses a novel combination of three evolutionary learning techniques, genetic programming (GP), evolutionary programming (EP), and genetic algorithms (GA) to evolve complete pattern recognition systems. The GP grows complex mathematical expressions that perform signal-to-signal transformations, EP optimizes convolution templates to process the results of these transformations, and the GA combines sets of feature detectors to form orthogonal features. A simple nearest neighbor classifier is used to classify the resulting features forming a complete pattern recognition system. This report provides a brief description of E-MORPH and presents recognition results for the problem of classifying high range resolution radar signatures. This problem is challenging because the data sets exhibit a large within class variation and poor separation between classes. The specific data set used in this experiment consists of 60 signatures of six airborne targets drawn from a  $1^\circ \times 10^\circ$  (azimuth x elevation) view window. The best recognition system evolved using EMORPH accurately classified 100% of the training signatures (6 targets x 5 samples = 30 signatures) and 90.0% of the signatures in an independent test set (6 targets x 5 samples = 30 signatures). This result is based on a preliminary experiment that did not involve tuning EMORPH's control parameters for this specific problem. This suggests that even better performance can be achieved in future experiments. The techniques used in E-MORPH are not tied to radar signals. The approach is generic and readily transitions to many different problems in automatic target recognition.

## HYBRID EVOLUTIONARY LEARNING SYSTEM

Mateen M. Rizki  
Associate Professor  
Department of Computer Science and Engineering  
Wright State University

### INTRODUCTION

The foundation of a robust pattern recognition system is the set of features used to distinguish among the given patterns. In many problems, the features are predetermined and the task is to build a system to extract the selected features and then classify the resultant measurements. In automatic target recognition problems, the identification of a set of robust, invariant features is complicated because the shape and orientation of the objects of interest are often not known *a priori*. As a result, a human expert is responsible of examining each problem to formulate an effective set of features and then build a system to perform the recognition task. An alternative to this labor intensive approach of building recognition systems has emerged in the past ten years that uses learning algorithms such as neural networks and genetic algorithms to automate the process of feature extraction. There are many advantages to the automated construction of recognition systems over techniques that rely solely on human expertise. Automated approaches are not problem specific. Consequently, once an automated system is developed, it can be readily applied to similar problems greatly reducing the time needed to solve new recognition problems. Automated systems are capable of producing solutions that are comparable to the customized solutions created by human experts, but the solutions formed by these systems are often non-intuitive and quite different from the solutions formed by human experts. In many applications, this is a drawback because it is not possible to describe how the solution is obtained. This is also a strength of the automated approach. Automated techniques are unbiased. The features selected to solve problems represent alternative designs based on the structural and statistical attributes of the data. The fact that different features are selected suggests that automated systems are capable of exploring different regions of the space of potential solutions.

Several automated target recognition systems exist that use evolutionary learning to extract features from raw data and perform classification [Rizki et al. 1993, 1994]. Early experiments with EMORPH, a system developed to evolve morphological algorithms, demonstrated that hybrid evolutionary learning systems are capable of generating pattern recognition systems to automatically perform feature extraction and classification from grey-scale images. In this system, a robust set of features is identified using a population of pattern recognition systems. Each system is composed of a collection of cooperative feature detectors and a classifier that evolves under the control of a user provided performance measure. The performance measure is tied to recognition accuracy, but additional constraints are included such as complexity measures to sculpt specific types of solutions. The recognition systems compete for survival based on their performance. Successful systems have a higher probability of survival and contribute more

information to future generations. The structural and statistical information gathered by each recognition system during the evolutionary process is passed to the next generation through a process of reproduction with variation. The most successful recognition systems are combined to form new recognition systems that are often superior to either parental unit. Two opposing forces operate in the evolutionary process: exploration and exploitation. By recombining successful solutions during reproduction, each generation contains recognition systems that are more capable of exploiting the performance measure and solving the recognition task. The reproductive process is imperfect, variations in the new recognition systems are created by mutating the structure of the feature detectors. Each new recognition system contains pieces of past successful designs with variations that explore alternative designs. The process of reproduction with variation and selection continues until the best recognition system in the population achieves a satisfactory level of performance.

This report describes experiments conducted using a modified version of EMORPH to evolve pattern recognition systems to classify high range resolution radar cross sections. This version of EMORPH blends three evolutionary learning paradigms: evolutionary programming [Fogel et al., 1966; Fogel, 1991], genetic algorithms [Holland, 1975; Goldberg, 1989], and genetic programming [Koza, 1992] to form a hybrid learning system. The system extracts features from a training set of radar cross sections, assembles cooperative sets of features, and forms a nearest neighbor classifier to label targets. A minimum amount of effort was devoted to tuning the EMORPH for the specific problem of classifying radar target, yet the evolved pattern recognition systems accurately classify an independent test set of radar cross sections.

## THE EVOLUTIONARY LEARNING SYSTEM

The overall design of the EMORPH generated recognition system is shown in Figure 1. A recognition system is composed of a feature extraction module and a classification module. The feature extraction module applies a set of feature detectors to each radar cross section to form a feature vector. The classifier then assigns a target label to each feature vector. A feature detector is composed of two components, a transformation and a cap. Transformations are networks of morphological and arithmetic operations that alter the signal in an attempt to enhance the most discriminating regions of the radar cross sections while suppressing noise. Caps are convolution kernels or templates composed of a collection of positive and negative Gaussian probes that are used to explore both the geometrical structure and contrast variation of the transformed signal. The convolution operator produces its strongest response when all of the positive and negative probe points align with similar regions in the signal. Consequently, when a cap produces a strong response, it indicates that geometry and contrast variation embodied in the cap also exists in the transformed signal. By adjusting the positions, values, and spread of the probe points, complex structural relationships are readily identified. The output of a detector set is a registered stack of processed radar cross sections. The set of caps present in a single recognition system is a registered set of convolution templates that serves as a 3D probe. This probe allows the recognition system to explore relationships within a single stack-plane (transformed

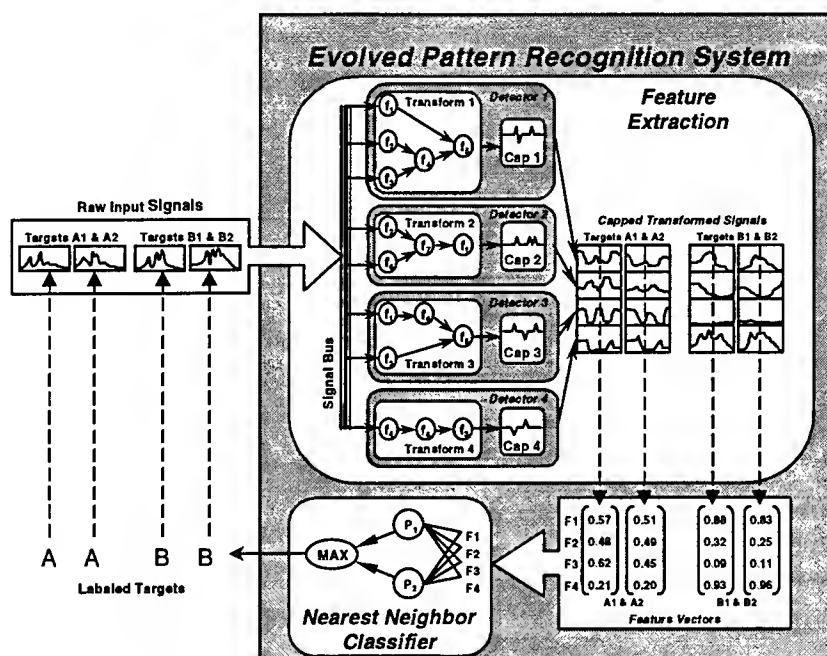


Figure 1. Overview of an evolved pattern recognition system.

view of the input signal) or across several stack-planes. By varying the parameters of each transform, the feature extraction module can decompose a signal into different spatial frequencies creating a pseudo-wavelet transformation. When this occurs, the 3D probe can exploit multiple resolution levels to selectively mask noisy high frequency spikes leaving only the most prominent structures for further analysis. When the full set of detectors is applied to a radar signal, a real valued feature vector is produced. Each detector contributes one component to the vector. By repeating this process for all the signals, a feature matrix is created that is used to form a nearest neighbor classifier for target classification.

The E-MORPH system forms feature detectors by creating a pool of signal transformation as shown in Figure 2. These transformations are evaluated using local performance measures that attempt to evaluate the information content of each transformed signal. The results of these transforms are capped using convolution templates and the results of the capped transforms are evaluated using a second local performance measure. Finally, capped transforms are selected to form a cooperative set of feature detectors that are evaluated by forming a nearest neighbor classifier to evaluate recognition accuracy. After each recognition system is assigned a performance measure, it competes for survival with other members of the population. The competition is organized as a tournament that ranks each recognition system based on its performance relative to the performance of other systems in the population. The size

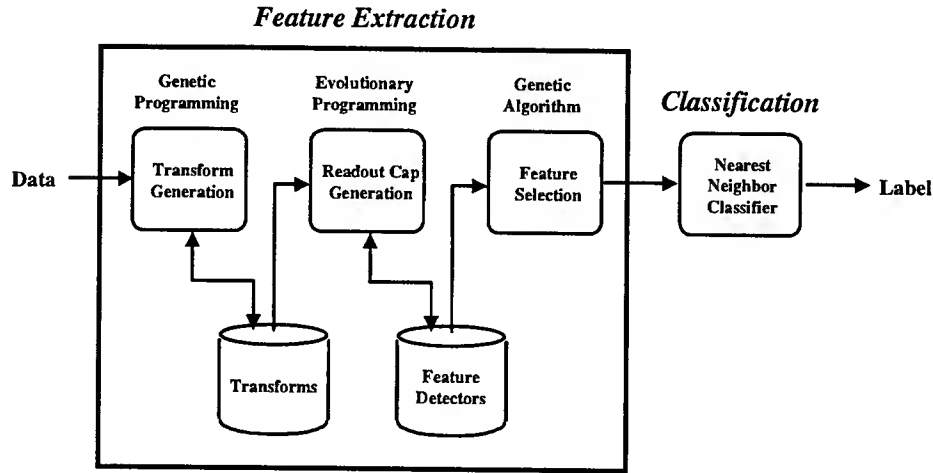


Figure 2. Overview of the EMORPH system.

of the tournament changes throughout the evolutionary process and is based on the average performance of the population as shown in Equation 1.

$$NC = \max \left( 1, M \cdot \frac{\sum_{i=1}^N pm_i}{N} \right) \quad \text{Equation 1}$$

In this Equation, NC is the number of competitors in each tournament, N is the population size, and M is a user imposed upper limit on the number of competitions ( $M \leq N$ ). Each recognition system must win as many conflicts as possible to increase its chance for survival. The number of competitions won or lost is calculated using Equation 2.

$$win_i = \sum_{k=1}^{NC} \left[ \left( \frac{pm_i}{pm_i + pm_{2 \cdot N \cdot U(0,1)}} \right) < U(0,1) \right] \quad \text{Equation 2}$$

In these local competitions, the chance of winning is proportional to the ratio of the performance measure of the recognition system and its competitor. For example, if a recognition system's performance ( $pm_i$ ) is high and a randomly selected competitor's performance ( $pm_{2 \cdot N \cdot U(0,1)}$ ) is low, then the probability that the ratio is greater than a value drawn from a uniformly distributed random variable  $U(0, 1)$  is also high. When the relationship shown in Equation 2 is satisfied, the recognition system wins the pairwise competition. Limiting the tournaments to a subset of the population reduces the possibility of premature convergence of the evolutionary process. When the average performance of the population is poor, the number of individuals in each tournament is small and a marginally better recognition system does not have the opportunity to dominate the population. The pairwise competition used within



each tournament tends to maintain a diverse population of recognition systems because marginal individuals always have a small probability of survival. The final selection for survival is based on a ranking of the number of conflicts won by each recognition system. The sets with the greatest number of victories survive to the next learning cycle.

E-MORPH uses three different techniques to alter the structure of the detector set contained in each recognition system. The position of the Gaussian points in the convolution templates within a detector set are varied using evolutionary programming [Fogel, 1991], the functional form of the transformation is modified using genetic programming [Koza, 1992], and the collection of detectors that form the basis of the feature extraction module are selected using a genetic algorithm [Holland, 1975]. These techniques are combined to exploit the strengths of each paradigm.

EMORPH uses genetic programming (GP) to grow signal transformations. Transformations are networks of morphological, arithmetic, and special operators that are represented as expression trees. Each expression performs a mathematical transformation of the input signal. The performance of a transformation is evaluated using a local performance measure that consists of a weighted sum of the total energy of the transformed signal, the number of peaks in the signal, the magnitude of the strongest peaks, and the distance between the strongest peaks. In addition, the performance is adjusted so that transforms producing similar effects receive a diminished score. The GP algorithm operates on a population of transformation. Parental units are selected from the base population using roulette wheel sampling where the probability of selection is proportional to the transform's performance measure. The transformations are represented as expression trees. The input patterns flow from the leaves of the tree through the operators to the root of the tree. The GP algorithm exchanges sub-trees between pairs of transformations. In Figure 3, transform one (dark grey) contains a root and two sub-trees labeled S1 and S2 while transform two (stippled grey) consists of a root and two different sub-trees labeled T1 and T2. Recombination forms two new transformations where the sub-trees S1 and T1 are exchanged in the offspring. In addition to exchanging information by recombination, sub-trees can be added, deleted, or replaced. Mixing the structure of expressions produces radical changes in the operation of the offspring transform. This disruptive process facilitates the search for new functions. The probability of each type of action is defined by the user. Usually, the probability of mutation (addition, deletion, replacement) is lower than the probability of recombination because recombination preserves larger pieces of the structure and therefore is slightly less disruptive than mutation.

EMORPH uses a combination of morphological and arithmetic operators as a basis for its functional transformations. Mathematical morphology is a technique for probing the structure of signals or images using set theoretic operations [Serra, 1982; Haralick et. al. 1987]. Each morphological operation is a signal-to-signal transformation that applies a probe-like pattern, referred to as a structuring element, to an input signal to produce an output signal. By selecting the correct algebraic form and structuring elements, specific objects can be isolated or enhanced, but finding the combination of operators and probes to perform a given task is difficult even for an experienced morphological

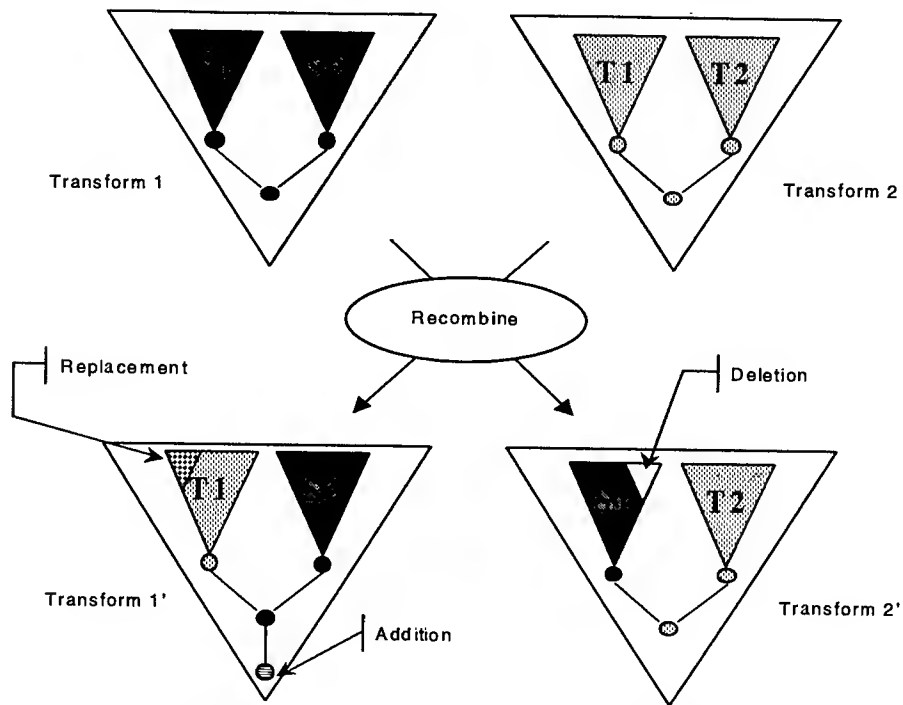
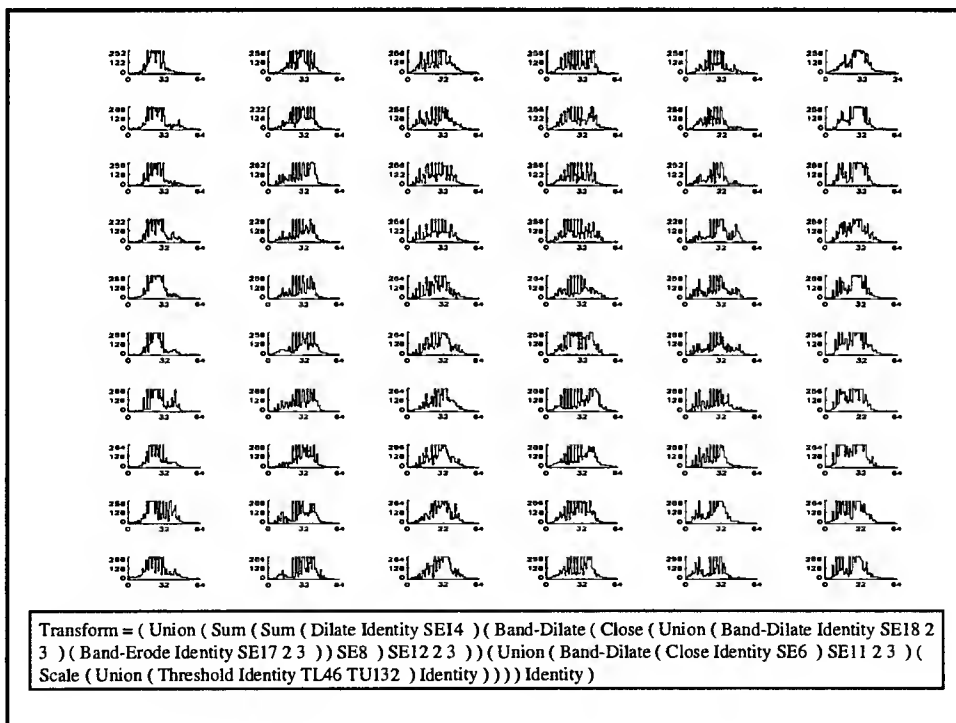
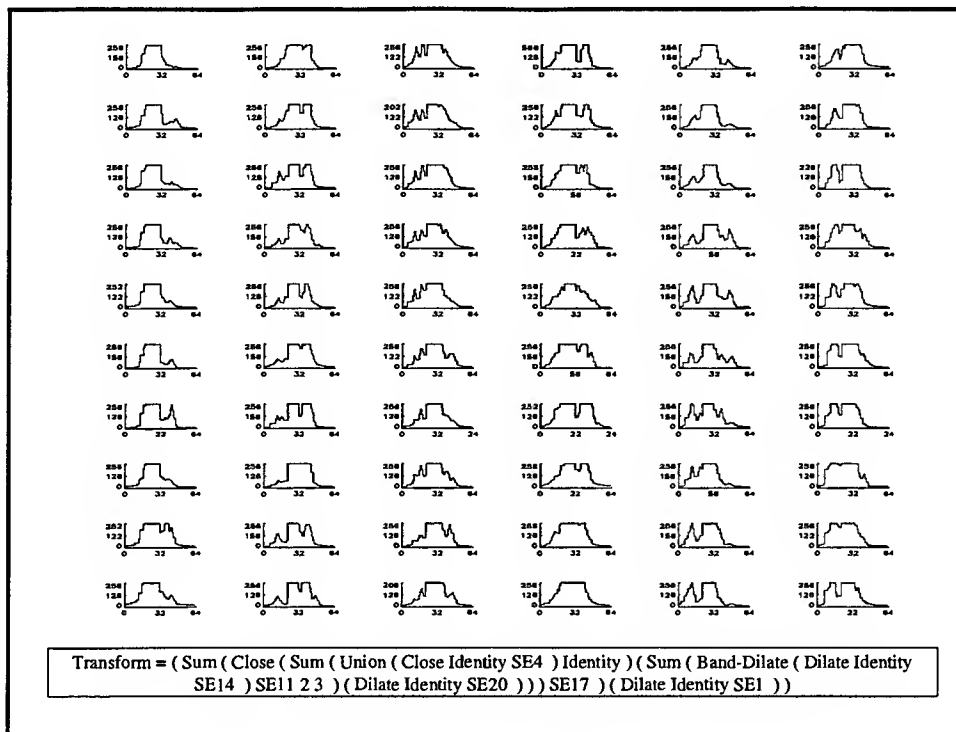


Figure 3. The GP process applied to transformation. Parental transformations are defined as expression trees. Sub-trees are recombined to form new transformations. During this process, some new operators are introduced (addition), some are removed (deletion), and some sub-trees are replaced with randomly generated sub-trees.

analyst. EMORPH solves this problem using GP to generate, evaluate, and select suitable morphological transformations to accomplish the desired classification. To begin the evolutionary process, the transformations that form the basis for the population of recognition systems are initialized with small randomly generated expression trees. Each node in the trees contains an operator drawn from the set: erosion, dilation, opening, closing, band-opening, band-closing, complement, addition, subtraction, minimum, maximum, and threshold. Most of the operators require some type of parameter. The morphological operators (erosion, dilation, opening, closing, band-opening, band-closing) use structuring elements that are selected at random from a standard library consisting of three basic shapes (e.g. 1-D cross section of a cone, a bar, a ball). A scale factor is also included to alter the size of the structuring element. Some of the arithmetic operators (minimum, maximum, threshold) also use a parameter to control the behavior of the operation. These parameters are selected from a uniformly distributed random variable ( $U(0,1)$ ). Detailed examples of morphological operations, library structuring elements, and the process of generating expressions are described by Zmuda et al. [1992]. When each expression is generated, it is applied to the input signals. If it produces an extreme effect (e.g., the output of the operation is a constant value), it is considered a lethal form and discarded. Transforms are generated until an acceptable pool is formed for the second stage of the process. Two sample transformations are shown in Figure 4, and the process of generating transformations is summarized in Figure 5.

The purpose of the EP algorithm is to systematically improve the position, type, and number of probe points in the



**Figure 4. Sample transformations.** The result of applying the transform to the training set is shown each bottom. The actual form of the transform is shown at the bottom of each box. The columns represent the radar signatures for six different targets. The row are variation in target elevation(-20, -18, ..., -12 degrees).

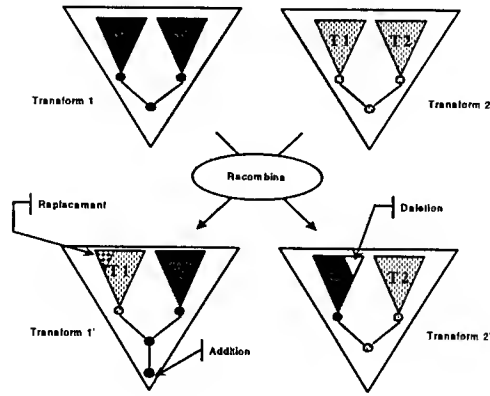


Figure 5. Overview of process used to evolve transformations.

convolution templates. This is accomplished using a controlled vibration of the position of the Gaussian points in each template followed by a series of random mutations that add and/or delete points (see Figure 6). The EP phase manipulate each recognition system independently. To begin, a member of the population of capped transformations is cloned to form an extended clonal population of  $C$  capped transformations. Each member of the clonal population then reproduces form an extended population. The caps in the extended population are subjected to random variations. The amount of variation is inversely proportional to the performance of the parental capped transformation and controlled by Equation 3. The value  $x_{j,k}$  is the central position of the  $k$ th Gaussian point in the  $j$ th

$$x_{j,k} = \max\left(\min\left(x_{j,k} + \left(\frac{X_{size_j}}{2} \cdot (1 - pm_i) \cdot N(0, 1)\right), X_{max_j}\right), X_{min_j}\right) \quad \text{Equation 3}$$

convolution template,  $X_{size}$  is the size of the template,  $X_{min}$  is the location of the left side of the template,  $X_{max}$  is the location of the right side of the template,  $(1 - pm_i)$  is the complement of the performance measure of the  $i$ th detector set, and  $N(0,1)$  is a normally distributed random variable with a mean of zero and a variance of one. To update a probe point's position, the mean of the random variable is set to the value of the initial position of the probe point and the

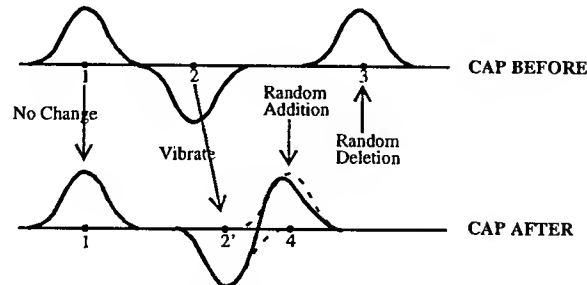


Figure 6. Evolving a convolution template.

variance is scaled to fall into the range from zero to half the template size. Using this technique, when the performance measure is low, the potential extent of variation in the position of a probe point is high. The potential for variation is reduced as the performance increases. If the performance reaches one, the potential for variation is zero and the template's point configuration is frozen. This approach to adjusting the structure of a template is similar to the process of simulated annealing where gradual improvements in the population performance shut down the process of random variation as a solution is formed.

The vibration process is only capable of adjusting the position of existing probe points. The second step of the EP phase is mutation that adds and/or deletes probe points to alter the complexity of the templates. Point mutation occurs immediately after the template points are vibrated. The amount of each type of mutation is controlled by a user selected probability. As a rule, if the detector set is initialized with a limited number of probe points, the probability of addition should be larger than the probability of deletion. This will bias the mutation rate toward addition and cause the detectors to grow in complexity.

In addition to the type and placement of the Gaussian points in a template, the variance (spread) of Gaussian probes change during the evolutionary process. The extent of each probe point is determined using an Equation 4. The

$$\sigma = \sigma_{min} + (1 - pm_i) \cdot (\sigma_{max} - \sigma_{min}) \quad \text{Equation 4}$$

limits on the spread of a single Gaussian point are set by the user to  $(\sigma_{min}, \sigma_{max})$ . The actual size of the probe point is then adjusted relative to the performance of the  $i$ th capped transform ( $pm_i$ ). If the resulting cap exhibits poor performance, the points increase in size to become less sensitive to the environment. If the cap is very accurate, the points become smaller and more sensitive to variations in the signals.

The decision to accept a mutated cap is based on a local performance measure. A value for the Fisher's Discriminant [Fisher, 1936] is calculated for the original capped detector and the mutated detector. This is a measure of the detector's ability to increase the separation between the means of the response for each class of target while simultaneously reducing the variance in the response for each class. If the mutated detector is more discriminating than the original detector, it replaces the parental unit. A few sample caps are shown in Figure 7 and the result of applying these caps to the transforms are shown in Figure 8.

After all the member of the clonal population reproduce, the C parental capped transform competes with the C offspring capped transforms for survival in a tournament. The top ranked C detectors are preserved and the evolutionary programming cycle begins again. After a fixed number of EP cycles, the performance of the best capped transform evolved during the EP phase is compare to the original parental capped transform. If the best evolved capped transform is more accurate than its parent, it replaces its parent in the base population. This process

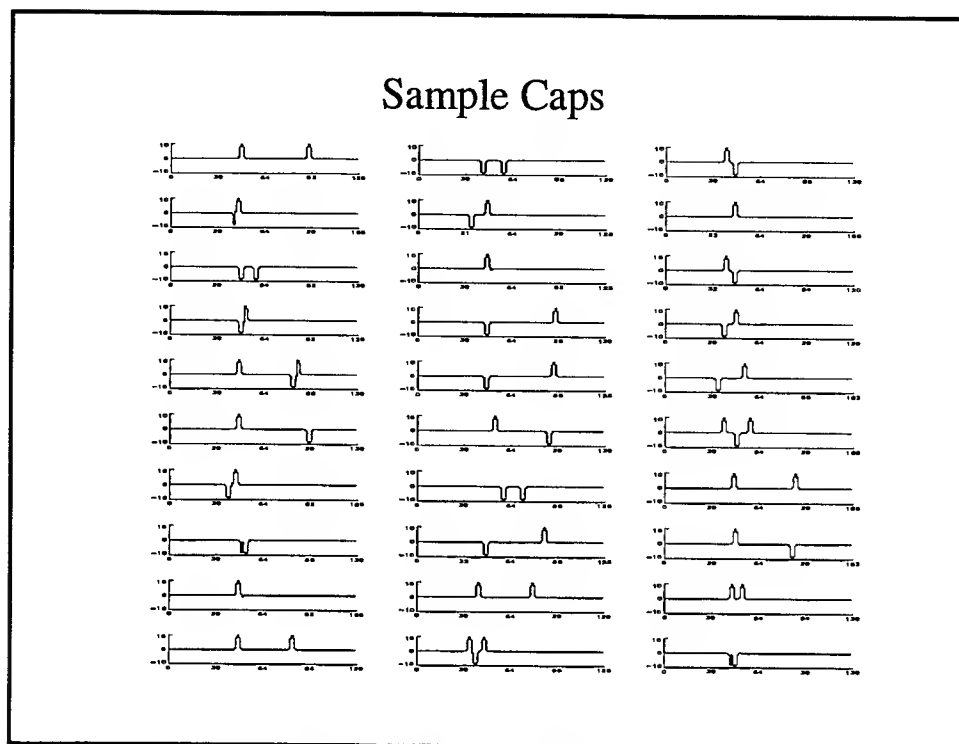


Figure 7. Sample convolution caps.

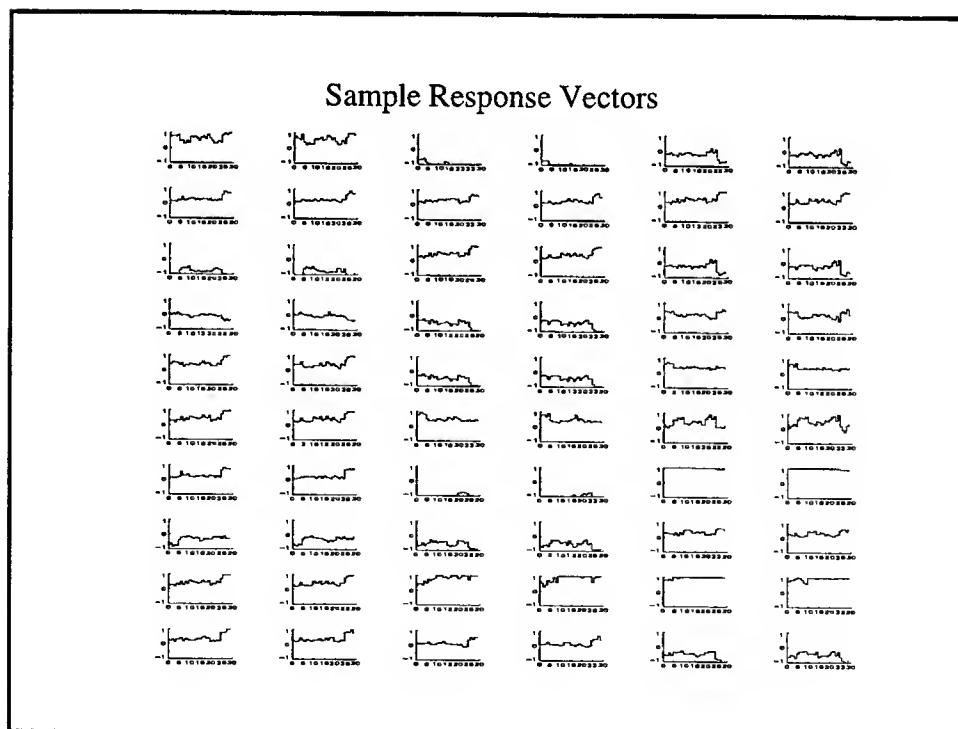


Figure 8. Response Vectors. The response vectors formed by applying the cap shown in the top-left corner of Figure 7 to the training data set.

### EP Algorithm

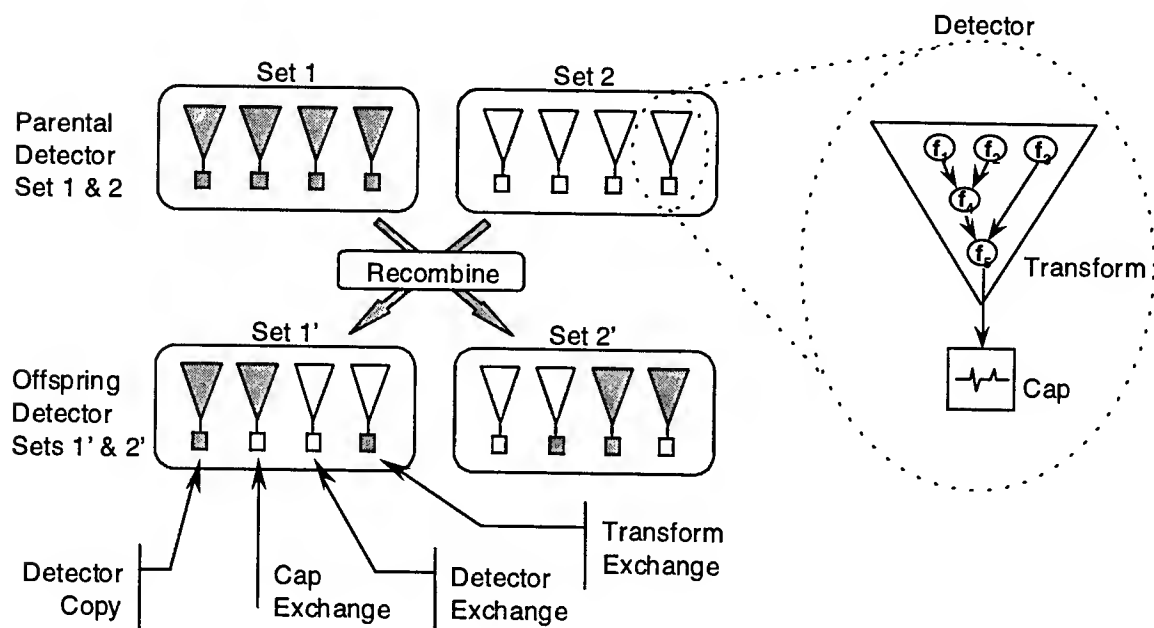
```
for each tranform Ti do
  Generate N random caps for Ti
  Evaluate capped version of the transform using Fisher's Discriminant
  for 1 to maxCycles do
    for each cap Cj do
      Produce an offspring cap by vibrating the point
      Mutate offspring (add/delete) points
      Evaluate the offspring
    endfor
    Perform tournament selection to rank the full population of caps
    Reduce population size to N based on ranking
  endfor
  Save the best cap for Ti
endfor
```

**Figure 9. EP Process used to generate convolution caps for transformations.**

is repeated for each member of the base population. When the EP phase terminates, the base population contains optimized sets of feature detectors consisting of convolution caps specifically tuned to the transforms produced by the GP phase. The process of evolving capped transformations is summarized in Figure 9.

EMORPH uses a genetic algorithm (GA) to form detector sets. The GA is responsible for recombining detector sets (see Figure 10). Parental units are selected from the base population using roulette wheel sampling where the probability of selection is proportional to the recognition system's accuracy. Once a pair of parents is selected, their detectors are exchanged using a uniform crossover. The detector set is analogous to a biological chromosome and the individual detectors are similar to genes. During crossover, each detector position in the parental set contributes some portion of each of its detectors to a pair of offspring detector sets. There is a 0.5 probability that the first parent places its information in the first offspring and the second parent places its detector in the second offspring. Similarly, there is a 0.5 probability that the first parent places its information in the second offspring and the second parent places its detector in the first offspring. As shown in Figure 10, a parental unit simply copies its whole detector (cap plus transform) into the selected offspring.

The GA phase begins with N detector sets and combines N/2 pairs of parental sets to form an extend population of 2N sets. Each member of the extended population is evaluated using the same procedure described for the EP phase. A tournament selection process is applied to rank the entire population and the N top-ranked detector sets are preserved

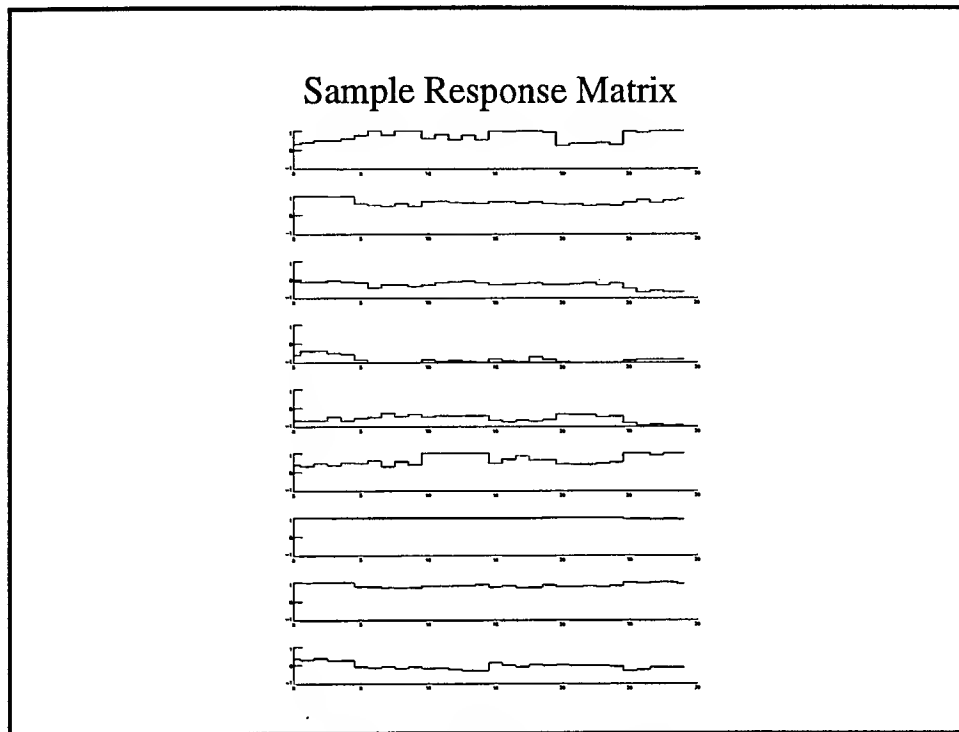


**Figure 10. Action of the GA process.** The GA takes pairs of detector sets and recombines components of each parental set to form pairs of offspring. This process can result in combinations of different caps on transforms or whole detectors being exchanged between sets.

for the next cycle of the GA algorithm. When the GA phase is complete, each detector set consists of combinations of transforms and caps that proved useful in the recognition process. A sample response matrix is shown in Figure 11. The rows represent the response of individual detectors. The x-axis is organized by target classes (1-5 is target one, 6-10 is target two, etc.). Notice how some detectors respond consistently to subsets of targets. The features defined by these detector readily form prototypes for a nearest neighbor classifier. The overall flow of the GA phase is outlined in Figure 12.

To summarize, EMORPH consist of four distinct stages: transform generation (form expression trees -- GP phase), detector formation (capping expressions -- EP phase), feature selection (forming response matrices -- GA phase), and finally, creating a nearest neighbor classifier to form the complete pattern recognition system. The user can set parameters to control the duration of each phase as well as control parameters to influence the behavior of each step of the process. For example, the user can increase the sensitivity of the individual detectors by increasing the number of passes through the EP phase relative to the number of passes through the GA phase. Alternatively, the user may elect to spend more computational resources adjusting the average complexity of the detector sets by increasing the number of passes through the GP phase. It is difficult to select an appropriate mixture of passes because the evolutionary learning process is dynamic. During the early stages of evolution, it is not likely that the complexity and number of detectors in the population is suitable for the recognition task. If the user arbitrarily increases the number of EP passes, the probe point density will increase to compensate for the lack of complexity in the transforms and





**Figure 11. Sample response matrix.**

### **GA Algorithm**

```

Form N random sets of detectors by sampling the pool produced by the EP process
Evaluate each detector set by forming the response matrix and classifying the training set
for 1 to maxCycles do
    Generate a selection vector based on classification accuracy
    for 1 to maxMatings do
        Select pairs of detector sets
        Perform uniform crossover on detector sets to form two offspring
        Mutate offspring by adding/deleting detectors
        Evaluate offspring detector sets
    endfor
    Perform tournament selection to rank the full population
    Reduce population size to N based on ranking
endfor

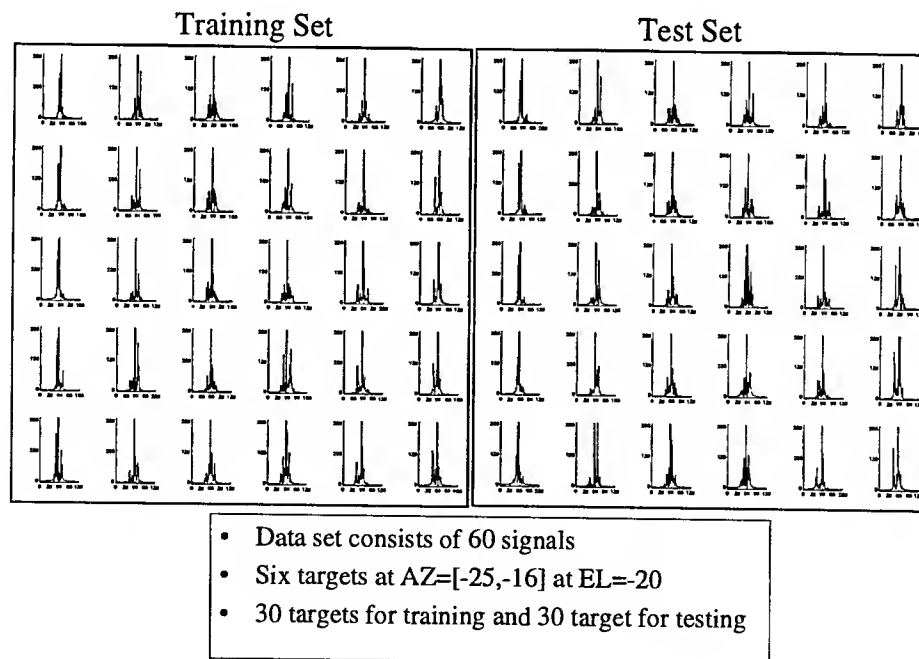
```

**Figure 12. GA Process used to select cooperative sets of feature detectors.**

limited number of detectors. This will produce customized solutions that tend to perform well on training sets and poorly on test sets. If the number of GP cycles is too large, the transform can become too complex to compensate for the inadequate distribution of probe points within each cap. A good compromise would be to implement an adaptive feedback mechanism to control the duration of each phase and adjust parameters relative to the contribution of each phase throughout the evolutionary process.

### EXPERIMENTAL DESIGN

To demonstrate how EMORPH generates pattern recognition systems, the results of a target recognition task in high range resolution radar are presented. Specifically, the problem is to classify a set of airborne targets from their radar cross sections. For this experiment, a sample of 60 radar signatures were extract from a large database of signals. Each radar signature is one view of a target at a specific azimuth and elevation. The selected data set contains six targets at azimuth  $25^\circ$  and elevations that range from  $-20^\circ$  to  $-11^\circ$  in increments of  $1^\circ$ . Thus, there are one ten samples of each target in the data set. These were divided into a training set of 30 radar signatures that contain 5 samples of each target and a test set of 30 cross sections that also contain 5 samples of each target. The data was not placed in the sets at random. The training set contains all targets with odd values of azimuth while the test set contains the remaining signatures. This amounts to placing every other signature in the view volume (azimuths x elevations) into one set and the remaining signatures into the other. Notice the signatures have been normalized into



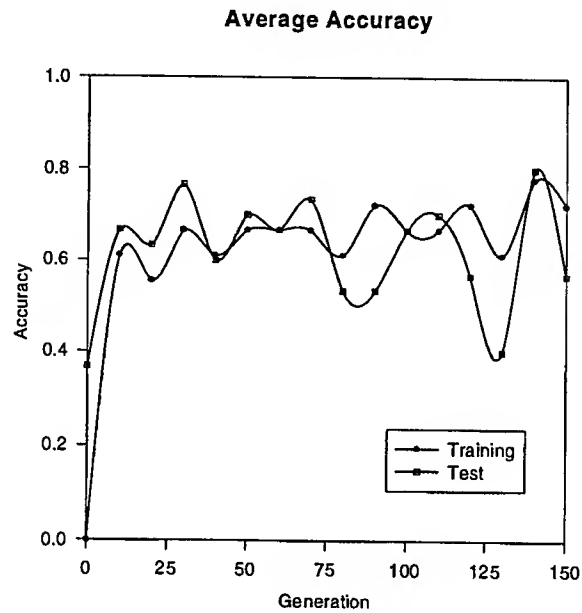
**Figure 13. Training and Test Data Sets.**

128 range bins with the maximum value (255) placed in bin 63. Looking down the column of data it is easy to see there are characteristic features in each target that persist through a few degrees of change in elevation, but then rapidly disappear. Also note the similarity in the signatures between targets making the classification task quite difficult.

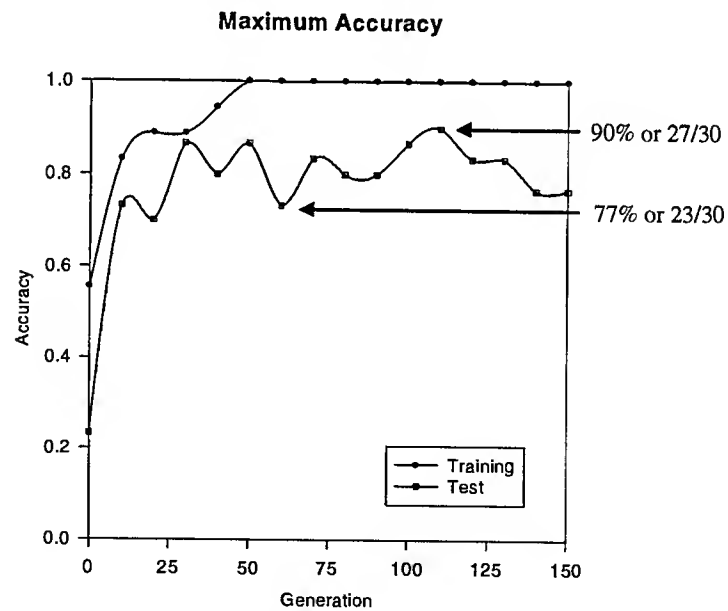
An E-MORPH learning cycle consists of thirty-five GP sub-cycles, followed by 35 EP sub-cycles, followed by 150 GA sub-cycles. To begin the GP phase, a population of 100 transforms was generated at random. Each transform was initialized with one to three operators. Each pass through the GP algorithm produced 100 offspring that were evaluated using the local performance measure. Then tournament selection was used to reduce the population back to 100 transforms. The EP phase was then applied to form a pool of feature detectors. The initial caps were generated with one or two randomly placed Gaussian points. The performance of each capped transform was evaluated by computing the Fisher discriminant for the response vector produced by applying the detector to the training set. A pass through the EP phase consists of processing each member of the base population of 100 transforms. Each member of the base population was used to produce 10 clones that are then mutated to produce an additional 10 recognition systems. This extended population of 20 caps was pruned back to 10 individual using tournament selection. The process was repeated five times and the best recognition system found competed to replace its ancestor in the base population. The GA phase started with the base population of 10 recognition systems. Copies of these base systems are mutated and recombined to create an extended population of 20 systems (10 parents + 10 offspring). The extended population was ranked using tournament selection and the top 10 systems were saved to start the next GA sub-cycle.

For the EP phase, the probabilities of vibration, addition, and deletion are 0.6, 0.3, 0.1 respectively. When a Gaussian point is added to a cap, there is a 0.67 probability that the point is positive and a 0.33 probability that the point is negative. The range of a Gaussian probe point is 4 to 12 range bins (i.e. pixels) and the maximum weight of a point is limited to the range of 1 to 3. In the GP phase there is a 0.3 probability that a transform is mutated and a 0.7 probability that a pair of transforms is recombined. If a set is selected to undergo mutation, there is a 0.2 probability that an individual transform passes to the offspring unchanged; a 0.7 probability that the transform is extended with a randomly selected operator, and a 0.1 probability that a random tree is added to the transform.

The average recognition accuracy for the population produced during the evolutionary learning process is shown in Figure 14. The performance is displayed at the end of each GA cycle. The average training set recognition accuracy rises from a low of 38% in generation 0 to 69% in generation 150. Notice, the curve is not monotonically increasing. This is because the uniform crossover used in the GA algorithm can extend and/or shrink detector sets producing disruptive effects. Also, tournament selection does not guarantee that the most accurate recognition systems will survive. The overall trend shown in this graph suggests that the selection process is locating increasingly accurate sets of features. Notice there is more variation in the test set performance than the training set performance. This is



**Figure 14. Average recognition accuracy for the training and test set.**

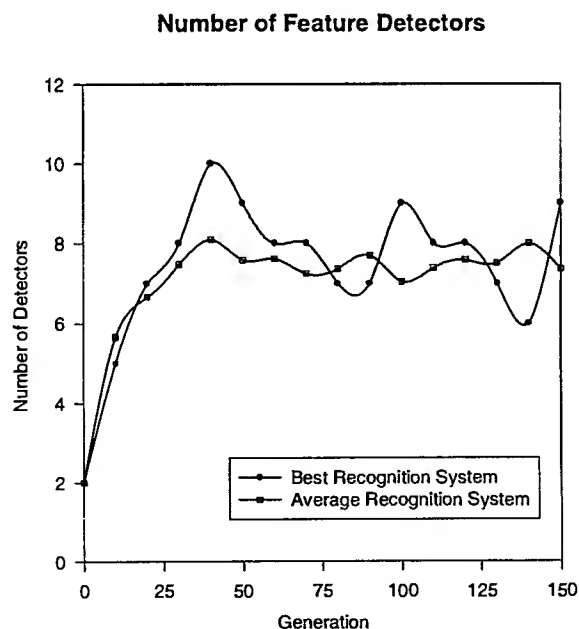


**Figure 15. Recognition accuracy of the best pattern recognition system for the training and test data sets.**

unfortunate, because it suggests that the features selected for the recognition systems are not generalizing. The performance of the best recognition system in each generation is shown in Figure 15. The accuracy of the best system rises from approximately 57% accuracy on the training and 22% accuracy on test data sets using 3 detectors to a maximum level of 100% on the training set and 90% on the independent test set. The best recognition system uses 9 detectors (Figure 16). The system appears to converge to rapidly producing a best training set score of 100% in generation 50. This sharply reduces the amount of exploration that can occur in future generations. This is partly due to the same population size used in this experiment. It is interesting to note that several recognition systems appear with test set accuracies approaching 100%, but these systems do not produce the top training scores. This suggests that a two-level training set might improve performance. One set would be used to form the recognition system and a second set would then be used to determine survival based on how well the system classified the secondary training set.

## DISCUSSION

E-MORPH successfully generated a pattern recognition system to classify high range resolution radar signatures. The evolved recognition system achieved a classification accuracy of 100% when applied to a training data set consisting of 30 radar signatures (5 samples of six targets) and 90% accuracy on an independent set of 30 additional signatures. The best recognition system contained nine feature detectors composed of primitive morphological and



**Figure 16. Number of feature detector used in the average and most accurate recognition systems.**

arithmetic operators capped by a special convolution template containing an evolved distribution of Gaussian-shaped probe points. The response of these detectors are processed by a simple nearest neighbor classifier that labels each signature. The use of morphological operators in the construction of primitive feature detectors allows EMORPH to evolve wavelet-like transformations that eliminate noise from the signatures and suppress information at various spatial frequencies to facilitate the process of classifying targets.

Although EMORPH achieves excellent recognition results, its performance can be improved. Inspection of the evolved feature detectors suggests that various redundant sub-expressions within the detector transformations can be eliminated to accelerate the evolutionary search process. This also implies that adjusting the library of operators and parameters used to grow feature detectors may improve both accuracy and the robustness of the evolved recognition systems. In addition, EMORPH's control parameters were not carefully tuned for this specific problem. Consequently, even better performance can be achieved in future experiments by adjusting the library of morphological operators, structuring elements, and distribution of the computation resource among the different phases of the evolutionary process.

The techniques used in EMORPH are not tied to radar signal processing. The approach is generic and can readily transition to many different problems in automatic target recognition. No single approach solves all problems in automatic target recognition, EMORPH represents one viable alternative. The solutions generated using our evolutionary learning algorithm are quite different than the solutions produced by human experts. This indicates that human experts may not be using all of the available information to develop robust pattern recognition systems. In future work, we hope to tune EMORPH, perform a more definitive set of experiments, and explore the possibility of combining human expertise with the evolutionary search process to access alternative designs. This hybrid approach to design may ultimately produce recognition systems with performance superior to any in use today.

#### ACKNOWLEDGMENT

I would like to express my appreciation to Dr. Louis Tamburino for serving as my laboratory focal point for the Summer Faculty Research Program. His total commitment to this research made my stay at Wright-Patterson Air Force Base a pleasure. I enjoyed his many helpful ideas and stimulating discussions over the years. I would also like to thank Dale Nelson and Jerry Covert for again allowing me the opportunity to participate in the summer program and providing a stimulating environment in which to work.

## REFERENCES

- Fisher, R. A., (1936). "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, 7, Part II, 179-188.
- Fogel, D. B. (1991). *System Identification Through Simulated Evolution: A Machine Learning Approach to Modeling*, Needham, MA: Ginn Press.
- Fogel, L. J., A. J. Owens, and M. J. Walsh (1966). *Artificial Intelligence Through Simulated Evolution*. New York, NY: John Wiley & Sons.
- Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*, Reading, MA: Addison-Wesley.
- Haralick, R. M., S. R. Sternburg, and X. Zhuang, (1987). "Image Analysis Using Mathematical Morphology", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-9:532-550.
- Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: The University of Michigan Press.
- Koza, J. R. (1992). *Genetic Programming: On the Programming of Computers By Means of Natural Selection*. Cambridge, MA: The MIT Press.
- Rizki, M. M., L. A. Tamburino, and M. A. Zmuda (1993) Evolving multi-resolution feature detectors. In *Proceedings of the Second Annual Conference on Evolutionary Learning*, eds. D.B. Fogel and W. Atmar, La Jolla, CA: Evolutionary Programming Society, 57-66.
- Rizki, M. M., L. A. Tamburino, and M. A. Zmuda (1994) E-MORPH: A two-phased learning system for evolving morphological classification systems. In *Proceedings of the Third Annual Conference on Evolutionary Learning*, eds. A. V. Sebald and L. J. Fogel, River Edge, NJ: World Scientific, 60-67.
- Serra, J., (1982). *Image Analysis and Mathematical Morphology*, London: Academic Press.
- Zmuda, M. A., M. M. Rizki, and L. A. Tamburino, (1992). Automatic generation of morphological sequences, In *SPIE Conference on Image Algebra and Morphological Image Processing III*, pp. 106- 118.

**AUTOMATED MODULAR FIXTURE PLANNING FOR VIRTUAL MATERIALS  
PROCESSING: GEOMETRIC ANALYSIS**

**Yiming (Kevin) Rong  
Associate Professor  
Manufacturing Systems Program**

**Southern Illinois University at Carbondale  
Carbondale, IL 62901-6603**

**Final Report for:  
Summer Faculty Research Extension Program  
Wright-Patterson Laboratory**

**Sponsored by:  
Air Force Office of Scientific Research  
Bolling Air Force Base, DC**

**and**

**Wright-Patterson Laboratory**

**December 1996**



# **AUTOMATED MODULAR FIXTURE PLANNING FOR VIRTUAL MATERIALS PROCESSING: GEOMETRIC ANALYSIS**

**Yiming (Kevin) Rong  
Associate Professor  
Manufacturing Systems Program  
Southern Illinois University at Carbondale**

## **Abstract**

Attendant Processes such as fixture and die design are often a necessary but time consuming and expensive component of a production cycle. Coupling such attendant processes to product design via feature-based CAD will lead to more responsive and affordable product design and redesign. In the context of on-going research in automating fixture configuration design, this report presents a fundamental study of automated fixture planning with a focus on geometric analysis. The initial conditions for modular fixture assembly are established together with needed relationships between fixture components and the workpiece to be analyzed. Of particular focus is the design of alternative locating points and components, together with example 3-D fixture designs.

# AUTOMATED MODULAR FIXTURE PLANNING FOR VIRTUAL MATERIALS PROCESSING: GEOMETRIC ANALYSIS

Yiming (Kevin) Rong

## 1. Introduction

Global competition has forced U.S. manufacturers to reduce production cycles and Product design agility. Generally, a manufacturing process is uncoupled and divided into: product design, process design (selection, routing and tooling), and assembly. Obvious and continual advances in computer-aided design (CAD), computer-aided process planning (CAPP) and computer-aided manufacturing (CAM) are enabling more multi-disciplinary design. However, computer-aided tooling (CAT), which is a critical part of process design and a bridge between CAD and CAM together with CAPP, has been least addressed and remains a missing link.

As a consequence of evolving CNC technology, specifically re-usable objects called features coupling shape and process (milling, drilling, etc.) to generate machine specific NC code, workpiece setup and associated fixturing has become the process bottleneck. To address this bottleneck, research and development of flexible fixturing, including modular fixturing technology, has received continued support. Modular fixture components enable a large number of configurations to be derived, disassembled and re-used. However, modular fixture design is a geometrically complex task and such complexity impedes widespread application of modular fixtures. Development of an automated modular fixture design system is needed to simplify process design of more affordable products. This research focuses on a geometric analysis for automated modular fixture planning which is inspired by several previous research in this area, especially a modular fixture synthesis algorithm [1] and an automated fixture configuration design methodology [2].

### Previous Research

Fixture design involves three steps: setup planning, fixture planning, and fixture configuration design [2]. Setup planning research has been addressed in the context of CAPP [3, 4, 5]. Seminal work in computer-aided fixture design (CAFD) focused on fixture planning:

- \* a method for automating fixture location and clamping [6];
- \* an algorithm for selection of locating/clamping positions providing maximum mechanical leverage [7];
- \* kinematic analysis based fixture planning [8, 9]; and
- \* rule-based systems to design modular fixtures for prismatic workpieces [10, 11].

But with respect to previous work on automating the configuration of workpiece fixtures, i.e., automated fixture configuration design (AFCD), little can be found. Fixture design depends upon critical locating and clamping points on workpiece surfaces, for which fixture components can be selected to hold the workpiece based on CAD graphic functions [12]. A 2-D modular fixture component placement algorithms has been developed [13]. In addition, a method for automating design of the configuration of T-slot based modular fixturing components has been developed [14]. A prototype AFCD system has been developed, including three core modules: fixture unit generation and selection module, fixture unit mount module, and interference checking module [2]. Assembly relationships between fixture components have also been defined and automatically established [15].

Nearly all the CAFD researchers admit that workpiece geometry is the pivotal factor in a successful CAFD system. Since the geometry of workpieces may vary greatly, many researchers in CAFD consider only regular workpieces, i.e., workpieces suitable for 3-2-1 locating method. There have been some attempts toward handling more complicated workpiece geometries as in reference [16]. However, their results are only applicable to some specific geometry, i.e., regular polygonal prisms.

### Review of Brost-Goldberg Algorithm

Recently, research in modular assembly based on geometric access and assembly analysis has gained considerable attention. Reference [1] presented a “complete” algorithm for synthesizing modular fixtures for polygonal workpieces and reference [17] explored the existence of modular fixture design solutions for a given fixture configuration model and a workpiece. Fixture foolproofing for polygonal workpieces was studied [18], and partially employed the approach in reference [1]. Reference [19] presented a framework on automatic design of 3-D fixtures and assembly pallets, but no detailed design methodology, procedure and results were provided.

In reference [1], an algorithm which is called the Brost-Goldberg algorithm was presented for synthesizing planar modular fixtures for polygonal workpieces. The basic assumptions were that a workpiece can be represented with a simple polygon, locators can be represented as circles with identical radius less than half the grid spacing, the fixturing configuration will be three circular locators and a clamp, the base plate is infinite, and all the contacts are frictionless. In addition to polygonal workpiece boundaries a set of geometric access constraints are provided as a list of polygons with clamp descriptions and a quality metric. The output of the algorithm includes the coordinates of the three locators, the clamp, and the translation and rotation of the workpiece relative to the base plate. The implementation of the algorithm is as follows per step 1:

1. The polygonal workpiece and geometric access constraints are transformed by extending the workpiece by the radius of the locators which are treated as ideal points (Figure 1) [1].
2. All candidate fixture designs are synthesized by enumerating the set of possible locator setups. The possible clamp locations are also found with each locator setup and clamp location specifies a unique fixture.
3. The set of candidate fixtures are then filtered to remove those that cause problems, i.e., collision. The survivors are then scored according to the quality metric.

In step 2 placement of three circular locators on the base plate are evaluated while translating and rotating the workpiece relative to the base plate. An algorithm was also presented to find all combinations of the three edges, where two of them may be identical,

on the polygon with a satisfaction of hole-alignment conditions with the base plate (Figure 2) [1]. For each set of locators and associated contact edges, consistent workpiece configurations or workpiece positions are calculated. All the possible clamp positions are then enumerated based upon the constraint analysis of the constructed force sphere.

The algorithm is called a "complete" algorithm for planer modular fixture design because it guarantees finding all possible planner fixture designs for a specific polygonal workpiece if they do exist. However, the major limitations of the method are:

1. Only polygonal workpieces are considered, i.e., no curved surfaces are allowed in the workpiece geometry. In reality, many fixture design cases include cylindrical surfaces, or circular arcs in 2-D representations.

2. Only circular locating pins with uniform radius' are considered in the algorithm. In each modular fixture system, there are some other types of locators available and widely used in fixture designs.

3. The algorithm only considers 2-D workpieces. In practice, it can be applied only for prismatic workpieces having small height, i.e., 3-D fixture design problem is a great challenge.

4. There are some criteria necessary for locating and clamping design in addition to geometric considerations including: locating error, accuracy relationship analysis, accessibility analysis, and other operational conditions.

5. Clamp location planning is weak without the consideration of friction forces, which needs to be further improved.

In this report, modifications and extensions to the modular fixture synthesis algorithm are presented with regard to limitations 1, 2, and 3 above. Discussion of limitations 4 and 5 will be presented in a separate report [20].

## 2. Geometric Conditions

A prismatic-workpiece is typically regarded as a 2-D workpiece including a set of edges such as line segments and arcs, which are candidate locating edges. The locating and clamping design problem becomes one of finding a group of three locating edge combinations. For an explicit expression, let us define the set of expanded boundary edges

of the workpiece P as:

$$EBE(P) = \{e_i \mid e_i \in \text{line segments}; i \in NE\} \quad (1)$$

where NE is the number of candidate edges.

All combinations of three edges, two of which may be identical, on the polygon are enumerated as:

$$\text{Triplets}(P) = \{(e_i, e_j, e_k) \mid e_i, e_j, e_k \in EBE(P), \exists (a, b) \subset (i, j, k), a \neq b\} \quad (2)$$

Locator centers are designed to contact with edge combinations  $ec = (e_i, e_j, e_k)$  (Triplets (P)). Without loss of generality, it can be assumed that  $e_i$  contacts with a locator  $L_1$  at the origin of the base plate lattice based on the assumption of an infinite base plate. By translating and rotating  $e_i$  about the origin,  $e_j$  sweeps out an annulus centered on the origin, with inner and outer diameter equal to the minimum and maximum distance between  $e_i$  and  $e_j$ . The position set of the locator contacting  $e_j$  should be within the swept annulus as

$$P_2(e_i, e_j) = \{p_2(x, y) \mid \min\text{-dist}(e_i, e_j) < \text{dist}(p_2, p_1) < \max\text{-dist}(e_i, e_j)\} \quad (3)$$

where  $p_1$  = origin of base plate lattice.

Each  $p_2$  is evaluated for selection as the second locator  $L_2$  in contact with  $e_j$ . If  $L_1$  contacts  $e_i$  and  $L_2$  contacts  $e_j$ , a third locator  $L_3$  in contact with  $e_k$  must be pairwise consistent with both  $e_i$  and  $e_j$ . The envelope containing the region swept by  $e_k$  maintaining contact with the first two locators can be easily determined by independently considering each pair as

$$P_3(e_i, e_j, e_k) = \{p_3(x, y) \mid p_3(x, y) \in P_2(e_k, e_i) \cap P_2(e_k, e_j)\}, \quad (4)$$

which is the same as presented in reference [1].

### Assembly Relationship Analysis

From the above discussion, it has been shown that determining the positions and orientations of modular fixture components can be simplified as finding geometric entities such as line segments or arcs on the workpiece passing ideal points on the base plate after moving (translating and rotating) the workpiece relative to the base plate.

As shown in Figure 3, the relative position between the workpiece and the base plate can be represented by the relation of the workpiece and the base plate coordinate systems which are expressed as  $X_w O Y_w$  and  $X_b O Y_b$  respectively. Basically, there are three locator-workpiece contact situations as shown in Figure 4:

- \* Line segment contacts with a circular locator.
- \* Arc contacts with a circular locator.
- \* Arc contacts with a line.

When a locating edge  $L_i$  on the workpiece is required to pass a point  $P_i$  on the base plate, i.e., the locator center needs to be aligned to a tapped (or pin) hole on the base plate,  $L_i$  can be expressed by

$$r_i x_w + s_i y_w + t_i = 0 \quad (5)$$

$P_i$  can be expressed as [2]

$$P_i: (x_{bi}, y_{bi}) \quad (6)$$

where  $\begin{cases} x_{bi} = 2Tu + T(v \bmod 2) \\ y_{bi} = Tv \end{cases}$ ,  $u, v = -N, -N+1, \dots, -2, -1, 0, 1, 2, \dots, N$ , and  $T$  is the

spacing increment between the taped (or pin) holes on the base plate.

The workpiece is assumed to be translated by  $(x, y)$  and rotated by  $\theta$  relative to the base plate. To simplify the calculation, an inverse transform is considered by holding the workpiece fixed, and moving the base plate by  $(-x, -y, -\theta)$ . Then  $P_i(x_{bi}, y_{bi})$  is transformed to

$$((x_{bi} - x) \cos \theta + (y_{bi} - y) \sin \theta, (y_{bi} - y) \cos \theta - (x_{bi} - x) \sin \theta). \quad (7)$$

Thus, the condition for the modular assembly can be described as:

$$r_i [(x_{bi} - x) \cos \theta + (y_{bi} - y) \sin \theta] + s_i [(y_{bi} - y) \cos \theta - (x_{bi} - x) \sin \theta] + t_i = 0 \quad (8).$$

For a specific workpiece, its geometry shape is fixed which means the equation of the line is fixed, i.e.,  $r_i$ ,  $s_i$  and  $t_i$  are constant. The assembly points are given, which means  $x_{bi}$  and  $y_{bi}$  are constant. There will be three equations to solve three unknowns  $x$ ,  $y$  and  $\theta$ .

If an circular locator contacts with an arc centered at  $O_i(u_0, v_0)$  with radius  $R$ , the arc can be represented as:

$$|P_i O_i| = R, \text{ or } (x_w - u_0)^2 + (y_w - v_0)^2 = R^2 \quad (9)$$

The contact equation will be:

$$(u_0 - (x_{bi} - x) \cos \theta - (y_{bi} - y) \sin \theta)^2 + (v_0 - (y_{bi} - y) \cos \theta + (x_{bi} - x) \sin \theta)^2 = R^2 \quad (10)$$

When an arc centered at  $O_i(u_0, v_0)$  with the radius  $R$  contacts a line-contact locator such as V-pad or half-Vee which has an incline edge  $AB$ , the third situation happens.

Assume  $\beta_{min}$ ,  $\beta_{max}$  are the extreme directional angles of  $P_i O_i$  that makes the arc still maintain contact with  $AB$  (Figure 4.c). Therefore,

$$\beta_{min1} = \beta_{min} + \theta, \quad \beta_{max1} = \beta_{max} + \theta$$

$$O_i(x_1, y_1) = (u_0 \cos \theta - v_0 \sin \theta + x, v_0 \cos \theta + u_0 \sin \theta + y)$$

Since distance( $O_i, AB$ ) =  $R$ , the fixturing condition becomes:

$$[r_i(u_0 \cos \theta - v_0 \sin \theta + x) + s_i(v_0 \cos \theta + u_0 \sin \theta + y) + t_i]^2 = R^2(r_i^2 + s_i^2) \quad (11)$$

where line  $AB$  in base plate coordinate system is represented as:

$$r_i x_b + s_i y_b + t_i = 0 \quad (12)$$

and  $P_i O_i$  should be within  $\beta_{min1}$  to  $\beta_{max1}$ .

Determining the position of a planner workpiece requires three parameters:  $x$  and  $y$  coordinates as well as the rotational angle  $\theta$  of the workpiece coordinate system. When the workpiece is placed into the fixture, it should contact with the three locators with three edges numbered  $j, k$  and  $l$ . Each contact will provide an equation concerning the workpiece location  $x, y$  and  $\theta$ . Eqs. 8 and 10 can be generally presented as:

$$G_i(x, y, \theta) = 0, \quad i = j, k \text{ and } l. \quad (13)$$

The contour of the workpiece can be represented by a group of differentiable functions in terms of workpiece coordinates  $(x, y, \theta)$  relative to the base plate because of the translation and rotation of the workpiece:

$$G_i(x, y, \theta) = 0, \quad i = 1, 2, \dots, n \quad (14)$$

where  $n$  represents the number of candidate locating points.

### Uniqueness of 2-D Solutions

Once the workpiece is positioned, the orientation should be unique. Solving the three-equation set (Eq. 13) may provide a solution, however, it is also possible that no solution



or infinite solutions are possible. The no-solution situation means the third locator has been chosen from an image-locus, i.e., all possible solutions exist in the pseudo-locus, but the pseudo-locus will also provide the position of the third locator which is not possible. In other situations, there are an infinite number of solutions, e.g., when all three edges are parallel or when the three contact normals meet at a single point (Figure 5). These cases should be discarded since they do not constrain the workpiece to a unique location.

In order to obtain the necessary conditions for a unique solution, assume that the workpiece can be positioned while contacting all three locators in a position  $(x_0, y_0, \theta_0)$  with a disturbance:

$$G_i(x + \Delta x, y + \Delta y, \theta + \Delta\theta) = 0, i = j, k, \text{ and } l, \text{ or}$$

$$G_i(x, y, \theta) + \frac{\partial G_i}{\partial x} \Delta x + \frac{\partial G_i}{\partial y} \Delta y + \frac{\partial G_i}{\partial \theta} \Delta\theta = 0 \quad (15)$$

For a stationary locating,

$$\frac{\partial G_i}{\partial x} \Delta x + \frac{\partial G_i}{\partial y} \Delta y + \frac{\partial G_i}{\partial \theta} \Delta\theta = 0, i = j, k, l. \quad (16)$$

Therefore, the condition for the equation set to have single solution is:

$$\begin{vmatrix} \frac{\partial G_j}{\partial x} & \frac{\partial G_j}{\partial y} & \frac{\partial G_j}{\partial \theta} \\ \frac{\partial G_k}{\partial x} & \frac{\partial G_k}{\partial y} & \frac{\partial G_k}{\partial \theta} \\ \frac{\partial G_l}{\partial x} & \frac{\partial G_l}{\partial y} & \frac{\partial G_l}{\partial \theta} \end{vmatrix} = 0 \quad (17)$$

For a valid solution, it is also important to consider the workpiece tolerances. When the geometric dimensions of the workpiece vary in a certain range, the locating contacts should be maintained. Similar analysis can be conducted, but the specifics will be presented in a forthcoming report [20]. Some other valuable discussion on similar problems can be found in reference [21].

### 3. Assembly Analysis

In this section, various locators and clamps are considered in fixture planning. In order to apply the fixture planning algorithm discussed in the previous section, the geometric analysis for workpiece boundary expansion should be performed for actual locators and clamps. Generally, there are two types of locating edges for 2-D workpiece geometry: line segments and arcs which may lie in either internal or external contours. Several locator types are used for side locating, including round locating pins (Figure 6.a), locating towers (Figure 6.b), adjustable stops (Figure 6.c), half-Vees (Figure 6.d), V-pads (Figure 6.e), round hole pins (Figure 6.f) and diamond hole pins (Figure 6.g).

If the locating edge is a line segment, a round locating pin, locating tower and adjustable stop may be used. For an arc segment, half-Vee and V-Pad are considered first. However, round locating pin, locating tower, and adjustable stop may be also used for arc edge contacts. Generally, locating a 2-D workpiece requires limiting three degrees of freedom (DOF): two translation and one rotational. Three line or arc edges, two of which may coincide, should be selected for locating purposes. Thus, a locator configuration should be considered for sundry combinations. Table 1 shows the possible locator configurations with assigned preference and provides criteria for preliminary selections of locators and clamps.

It was shown in reference [17] that the three circular locating pin configuration were not universal for arbitrary 2-D workpiece. Indeed, there exist some workpieces which can not be fixtured using this configuration, and therein, the type of locator may be changed. An alternative may involve the use of adjustable stops with adjustable contacting lengths. The distance from the contact point to the locator center may be larger than half of the base plate grid distance, which may greatly improve the locating capability.

Locating geometric analysis is based on the geometric constraints imposed on the workpiece and locator position. Here locators must maintain contact with specific locating edges on the workpiece. The modular fixture assembly requires the locators to be assembled through holes in the base plate. For the 2-D situation, the assembly process is to find the suitable assembly holes in the base plate which can locate the workpiece. Following are several cases on how to find possible locator positions.

### Case 1: Locating with locating tower and adjustable stop

Locators used for line segments are first discussed, such as locating tower and adjustable stop (locator b and c in Figure 6), as shown in Figure 7. Locating towers can be treated as smaller circular locators whose radius  $r$  is

$$r = \text{distance (locator center, locating edge)} \quad (18)$$

However, it should be noted that for locating towers, the possible contact region between the locating tower and locating edge should be reviewed to ensure the functional stability of the locating tower (Figure 8):

$$L_e = L - d \quad (19)$$

where  $L_e$  is the effective locating edge length,  $L$  is original length of the locating edge, and  $d$  is the length of the locating surface.

The adjustable stop can be treated as a circular locator with a radius  $r$  as a variable min-acting-distance  $< r < \text{max-acting-distance}$  (20)

Using such a geometric representation, input geometry transformation may be used to perform geometric analysis by expanding the corresponding locating edges by the equivalent radius.

### Case 2: Locating with hole-pins

If the locator configuration employs circular round-pin or diamond-pin to locate with small internal holes, it is easy to do assembly analysis since the center of hole and hole pin should be aligned. As shown in Figure 9.a, the first step of assembly is to align the diamond-pin with a hole on the base plate. Then the workpiece has only one rotational DOF to find the suitable assembly holes for other locating edges. Generally, adjustable locators may be used to ensure the availability of assembly holes for the other two locating edges. The round-pin application is shown in Figure 9.b.

If two small holes are employed, either the distance between the two holes has to be standard as

$$O_1, O_2 = k T \quad (21)$$

where  $O_1, O_2$  are centers of two holes and  $T$  is the base plate grid distance. If this condition is not valid, adjustable-bar support should be used near the bottom of the workpiece for one hole locator to ensure the assembly of the hole locator, which then becomes a 3-D locating problem (Figure 10).

### Case 3: Arc segment locating using circular locators

When the locating edge is an arc, input geometry transformation can also be used by expanding the arc through the equivalent radius of the locator in the direction of external normal. It is applicable to both external and internal arcs. The locus analysis is almost the same as those presented for line segment situation. The major difference lies in calculating the workpiece location and orientation.

As described in section 2, when the first circular locator is placed in the base plate origin, by translating and rotating  $e_i$  about the origin,  $e_j$  sweeps out an annulus centered on the origin, with inner and outer diameter equal to the minimum and maximum distances between  $e_i$  and  $e_j$ . The position set of the locator contacting  $e_j$  should be within the swept annulus as

$$P_2(e_i, e_j) = \{p_2(x, y) \mid \min\text{-dist}(e_i, e_j) < \text{dist}(p_2, p_1) < \max\text{-dist}(e_i, e_j)\} \quad (22)$$

where  $p_1$  = origin of base plate lattice.

It should be noted that the  $e_i$  and  $e_j$  could be either line or arc segments when using circular locators. However, in the case of applying other types of locators with arc edges, such as V-pad and half-Vee, the way to find locator positions with hole alignment relationships needs to be further studied.

### Case 4: Locating with V-pad

In Table 1, when the locating triplet is composed of one line segment  $e_2$  and one external arc  $e_1$ , the recommended locator configuration is using one V-pad and one circular locating pin. As distinguished from circular locating pins, assembling a V-pad requires two locating holes in base plate instead one, and the orientation of the V-pad can not be arbitrary and must have four perpendicular orientations.

As shown in Figure 11, a V-pad is placed around the origin of the base plate and oriented in one of the four possible perpendicular orientations. The center of the locating arc  $O_1$  as well as the contacting points between V-pad and the workpiece are then determined. The position of the circular locating pin may be found through rotating the workpiece while maintaining a 2-point contact with the external arc  $e_1$  with the V-pad. The locus of the round locating pin is a part of the annulus centered in the fixed locating arc center whose inner and outer diameters of the annulus are the minimum and maximum distance between the arc center and the line segment  $e_2$ :

$$P_2(O_1, e_1, e_2) = \{p_2(x, y) \mid \min\text{-dist}(O_1, e_2) < O_1p_2 < \max\text{-dist}(O_1, e_2)\} \quad (23)$$

The angle scope of the partial annulus is determined by the possible rotation angle of the locating arc about the V-pad without loss of contact.

$$\alpha_{\min} - 90^\circ + \beta < \text{angle} < \alpha_{\max} - 90^\circ - \beta \quad (24)$$

where  $\alpha_{\min}$  is the minimum angle between  $e_1$  and  $e_2$  with reference  $O_1$ ;  $\alpha_{\max}$  is the maximum angle between  $e_1$  and  $e_2$  with reference  $O_1$ ; and  $\beta = 45^\circ$  for  $90^\circ$  V-pad or  $30^\circ$  for  $120^\circ$  V-pad.

#### Case 5: Locating with half-Vee

A locating configuration may require using one half-Vee (or other line-contact locators) for an arc segment. The assembly character of half-Vee locators is more complicated than a V-pad. The shape of a half-Vee is shown in Figure 12.a. There are three locating holes in one half-Vee. When assembling, two holes in the half-Vees are needed to be accurately aligned with two locating holes in base plate. There are only four possible directions for the half-Vee when assembled to the base plate. In this report, the two locating holes  $VLH_1$  and  $VLH_2$  with equal distance to the oblique edge are analyzed. Other half-Vee shapes should be addressed via the same method.

First, a half-Vee is placed in a specific position on the base plate by aligning  $VLH_1$  and  $VLH_2$  with two locating holes  $BLH_1$  and  $BLH_2$  centered at  $H_1$  and  $H_2$  in the base plate. When the given arc ( $e_1$ ) centered at  $O_1$  maintains contact with the half-Vee, it can be transformed as growing the given arc by  $r$  (Figure 12.a). The contact situation can be thought as an arc rolling over a line segment.

The second locator position relative to the second locating edge ( $e_2$ ) can be found using geometric locus analysis. When  $e_1$  maintains a one-point contact with a half-Vee, the workpiece can translate and rotate, and  $e_2$  sweeps out which may be confined to a geometry centered at the connection line of the two locating holes in the half-Vee. The geometry is derived by the swept partial annulus when the arc contacting the different position on the half-Vee simply rotates without slip. The locus may be further refined by considering angle limitations of arc rotation (Figure 12.b). Generally, the swept geometry by  $e_2$  can be defined as a ribbon by satisfying such conditions:

- 1) The locus geometry is relative to a reference line segment  $H_1H_2$  where:

$$\text{distance}(O_1, H_1H_2) = R_e \quad (25)$$

where  $R_e$  is the expanded radius of  $e_1$  and in the outer normal direction of  $H_1H_2$ .

- 2) The two limit line segments are determined by off-setting  $H_1H_2$  through

$$\begin{aligned} \text{dist1} &= \text{maximum-distance-refer-to-} H_1H_2 (O_1, e_2) \\ \text{dist2} &= \text{minimum-distance-refer-to-} H_1H_2 (O_1, e_2) \end{aligned} \quad (26)$$

which means the distance in the direction perpendicular to  $H_1H_2$  (Figure 12.c).

If the second locator is designed to be a circular locating pin, the position of the locator may be chosen among the generated locus. If the second locator is designed to be another half-Vee, the position of the second half-Vee may be found through a similar assembly analysis. Noting that the second half-Vee can be transformed to an ideal line segment, positioning of the second half-Vee is required to find the position of the line segment. The line covering the line segment should be first determined and then the relative line segment may be determined such that the line segment contacts the workpiece. Any line intersecting the generated locus may be a candidate. For the third locator placement, i.e., locating edge  $e_3$ , the position can be found by considering the intersection of the locus of two pairs:  $e_1$  with  $e_3$ , and  $e_2$  with  $e_3$ . The intersection can cover the swept geometry by maintaining contact with  $e_1$  and  $e_2$  with the workpiece. Figure 12.d shows the swept region of possible positions for the third locator, which needs to be further constrained by the feasible rotation angles of the workpiece relative to the half-Vees when they are in contact.

#### 4. 3-D Fixture Configurations

2-D fixture planning as discussed above is limited to prismatic workpieces where the height of the workpiece is relatively small. The vast majority of workpieces are three-dimensional, and therein it is desirable to extend the above 2-D strategies. Such a fixture configuration design system has been developed, where when fixturing points are specified, fixturing units can be automatically generated [2]. In this section, a 3-D automated modular fixture planning procedure is presented followed by 3-D assembly analysis.

#### 3-D automated modular fixture planning procedure

Prior to fixture planning, the orientation of the workpiece relative to the base plate as well as machining surfaces in each setup must be determined in setup planning. First, although the workpiece geometry could be very complex, only four kinds of surfaces need be considered for locating purposes: planes parallel to the base plate (surface type A), planes perpendicular to the base plate (surface type B), cylindrical surfaces with an axis parallel to the base plate (surface type C), and cylindrical surfaces with an axis perpendicular to the base plate (surface type D). A 3-D automated modular fixture planner is outlined in Figure 13.

#### A. Determination of candidate locating surface set

The first step in using a 3-D automated modular fixture planning procedure is to find all candidate locating surfaces based on the above 'four-kinds-of-surfaces' assumption. The candidate locating surfaces can be obtained by retrieving the CAD model of the workpiece. The candidate locating surface set can be further refined if we assume that locating can be divided into two types: horizontal and vertical locating. Surfaces of type B and type D can be used for horizontal locating. Surfaces of type A and type C can be used for vertical locating. For vertical locating, those planes whose external normal is opposite to the base plate are discarded.

### B. Locating surface group selection

The next step is to select horizontal locating surfaces and vertical locating surfaces from the candidate locating surface set. Generally, three surfaces for each locating purpose should be selected as a group. The three vertical locating surfaces could be reduced to a singular surface. The three horizontal locating surfaces could be reduced to two surfaces with one surface be chosen twice. The locating surface groups are selected by considering accuracy relationships, geometric accessibility, operational conditions. A priority index may be generated for each locating surface group so that the surface group with the highest priority will be processed first. If later this strategy fails to provide a reasonable fixture plan, the surface group with the next highest priority index is chosen until one reasonable fixture plan is generated.

### C. Horizontal locating

The third step involves horizontal locating. Horizontal locating surface groups have been chosen in the second step. Considering each side as a locating surface, one locating unit (which usually consists of one locator and several supporting components) is constructed by using the automated fixture configuration design functions [2]. When the height of the locating points are approximately determined, e.g., the half-height position of the side locating surfaces, the locating units for each side locating surface are generated with assembly relationships between fixture components of the units. The assembly analysis is then performed to place these locating units on the base plates.

Generally, the position of a 3-D workpiece is determined by six parameters: three translation parameters ( $x$ ,  $y$  and  $z$ ) and three rotational parameters ( $\alpha$ ,  $\beta$ ,  $\gamma$ ) about  $x$ ,  $y$ , and  $z$  axes. Since the workpiece should maintain relative orientation to the base plate, the rotational parameters about  $x$  and  $y$  axes,  $\alpha$  and  $\beta$ , are fixed. After all the side locating units are placed, three position parameters ( $x$ ,  $y$ , and  $z$ ) will be determined. The parameter  $z$  will be determined by the clearance requirement between the workpiece and the base plate.



#### D. Vertical locating

In the vertical locating, the locators are first chosen by considering the types and surface finish of vertical locating surfaces. Similarly, the vertical locating units are generated by applying the automated fixture configuration design functions.

#### E. Clamping design

In clamping design, The number and type of clamps employed should be first decided base on workpiece stability analysis and operational rules. All possible clamping faces are then collected into a set. A combination of several candidate clamping surfaces are then selected. Assembly analysis is performed to place the clamps on the base plate given the assembly character of the clamps. Detailed analysis and discussion of clamp planning can be found in reference [20].

It should be noted that automated modular fixture planning for 3-D workpieces is very complicated. This design methodology only provides a framework for fundamental analyses of 3-D automated modular fixture planning.

#### 3-D Modular Assembly Analysis

Modular assembly analysis is the focus of this report, where in modular assembly analysis for 2-D situations is expanded to 3-D. In 3D situations, locating units instead of locators are the major concerns when conducting the assembling analysis. Figure 14 shows a sketch of locating units. A locating unit typically consists of a locator on the top and several supporting components. Below, only horizontal locating units are discussed since the assembly of vertical locating units is relatively easy. The side locating units are divided into two categories based on the characters of their locators: direction-fixed, and direction-variable.

When a workpiece maintains contact with an edge bar, the contact direction is fixed. If the locator is a round locating pin, locating tower or adjustable stop, the contact direction of the locator can change randomly corresponding to the locating surfaces on the workpiece. Placing the direction-fixed locating units will pose additional constraints on the direction of the side locating surfaces. In other words, two direction-fixed locating

units may conflict if their locating directions are not compatible. However, using a direction-fixed locating unit will also simplify the assembly process because of the assembly constraints. Direction-variable locating units are often more flexible. Direction-variable locating units will be discussed below.

Given a locating unit, the bottom component is connected with the base plate. Generally, the bottom component may use two locating holes to accurately determine the position and orientation of the bottom component. If two locating holes are needed, the placement of the locating unit can have only four directions parallel to the base plate symmetrical axes. The other important component in the unit is the locator which contacts with the workpiece. When the locating unit is generated, all the components in the locating unit are determined and their relative positions are also determined [2]. Therefore, the relative position of the locator to the bottom component can be derived which is very important to assembly analysis.

In 3D situations it is assumed that there are three generated side locating units: SLU1, SLU2, and SLU3 which are designed to contact with the three side locating surfaces: S1, S2 and S3. First, the 3-D workpiece is projected onto the base plate and become a 2-D geometry. Since S1, S2 and S3 are planes or cylindrical surfaces perpendicular to the base plate, three segments of lines or arcs are achieved with respect to the three side locating surfaces. They are then expanded by the radius of each locator respectively to get three segments of lines or arcs (s1, s2 and s3) and the locators can be reduced to ideal points (Figure 15). SLU1 is placed around the origin of the base plate and locator 1 is also positioned. Thus, s1 should maintain contact with locator 1, while s1 can rotate and slip. S2 sweeps out an annulus centered at locator 1 just like the 2-D situation (Figure 16). The position of SLU2 can be determined by transforming all the possible placement origins of bottom components by the x, y offsets of the locator which may have four directions. All possible transformed placement origins falling inside the swept annulus will be suitable as candidate SLU2 locations. In the same way, SLU3 can be positioned by considering s3 pairwise consistent with s1 and s2.

When all side locating units are placed, their positions are sent to another module to calculate the x, y translation position and  $\gamma$  rotational position.

## 5. Examples and Summary

A geometric analysis for automated fixture planning has been presented, which is an expansion of previous research on automated fixture configuration design and 2-D geometric synthesis. Cylindrical surfaces, different types of locating components, and 3-D fixture configurations have been considered in the analysis. Figure 17 shows two examples of fixture designs resulting from the fixture planning and fixture configuration design. A comprehensive automated fixture planning and configuration design system is under development where analyses of locating accuracy, geometric accessibility, clamp planning, and fixture design stability are included.

## References

1. R. C. Brost and K. Y. Goldberg, "A complete algorithm for designing planar fixtures using modular components", IEEE Transactions on Robots and Automation, Vol. 12, No. 1, 1996, pp. 31-46.
2. Y. Rong and Y. Bai, "Automated Generation of Fixture Configuration Design", ASME Transactions, J. of Manufacturing Science and Engineering, Vol. 118, No. 4, 1996.
3. A. Joneja and T. C. Chang, "A Generalized Framework for Automatic Planning of Fixture Configuration," *Advances in Manufacturing Systems Engineering*, ASME WAM, San Francisco, CA, Dec. 10-15, 1989, pp. 17-28.
4. C. H. Chang, "Computer-Assisted Fixture Planning for Machining Processes," Manufacturing Review, Vol. 5, No. 1, 1992, pp. 15-28.
5. P. M. Ferreira and C. R. Liu, "Generation of Workpiece Orientations for Machining Using a Rule-based System," Int. J. of Robotics and CIMS, Vol. 5, 1988.
6. Y. C. Chou, V. Chandru and M. M. Barash, "A Mathematical Approach to Automatic Configuration of Machining Fixtures: Analysis and Synthesis," ASME Transactions, J. of Engr. for Industry, Vol. 111, 1989, pp. 299-306.
7. E. C. De Meter, "Selection of Fixture Configuration for the Maximization of Mechanical Leverage," *Manufacturing Science and Engineering*, ASME WAM, New Orleans, LA, Nov. 28-Dec. 2, 1993, PED-Vol. 4, pp. 491-506.
8. R. J. Menassa and W. DeVries, "A Design Synthesis and Optimization Method for Fixtures with Compliant Elements," *Advances in Integrated Product Design and Manuf.*, ASME WAM, PED-Vol. 47, Dallas, TX, Nov. 25-30, 1990, pp. 203-218.
9. M. Mani and W. R. D. Wilson, "Automated Design of Workholding Fixtures using Kinematic Constraint Synthesis," *16th NAMRC*, 1988, pp. 437-444.
10. A. Markus, E. Markusek, J. Farkas and J. Filemon, "Fixture Design Using Prolog: an Expert System," Int. J. of Robotics and CIMS, Vol. 1, No. 2, 1984, pp. 167-172.

11. D. T. Pham and A. de Sam Lazaro, "AUTOFIX - an Expert CAD System for Jig and Fixtures," Int. J. of Machine Tools & Manufacture, Vol. 30, No. 3, 1990, pp. 403-411.
12. R. Sakal and J. G. Chow, "A semigenerative Computer-aided Fixture Design System using Autocad and CAD Fixturing Database," *Computer-aided Production Engineering*, Cookeville, TN, Aug. 13-14, 1991, pp. 461-458.
13. A. J. C. Trappey, C. S. Su and S. H. Huang, "Methodology for Location and Orientation of Modular Fixtures," *Manufacturing Science and Engineering*, ASME WAM, New Orleans, LA, Nov. 28-Dec. 2, 1993, PED-Vol. 64, pp. 333-342.
14. K. Whybrew and B. K. A. Ngoi, "Computer-aided Design of Modular Fixture Assembly," Int. J. of Adv. Manuf. Tech., Vol. 7, 1990, pp. 267-276.
15. Y. Bai and Y. Rong, "Modular fixture element modeling and assembly relationship analysis for automated fixture configuration design", Special Issue on *Rapid Prototyping & Reverse Engr., J. of Engineering Design and Automation*, 1996.
16. B. Nnaji, S. Alladin and P. Lyu, "A framework for a rule-based expert fixturing system for face milling planar surfaces on a CAD system using flexible fixtures", Journal of Manufacturing Systems, Vol. 7, No. 3.
17. Y. Zhuang, K. Goldberg and Y. Wong, "On the existence of modular fixtures", *IEEE International Conference on Robotics and Automation*, May 1994.
18. K. Penev and A. Requicha, "Fixture foolproofing for polygonal parts", *IEEE Int. Symp. on Assembly and Task Planning*, Pittsburgh, PA, August 10-11, 1995.
19. R. C. Brost and R. R. Peters, "Automatic design of 3-d fixtures and assembly pallets", *IEEE International Conference on Robotics and Automation*, 1996.
20. Y. Wu, Y. Rong, W. Ma, and S. LeClair, "Automated Modular Fixture planning: Geometric Analysis," (in preparation).
21. H. Asada and A. By, "Kinematic analysis of workpart fixturing for flexible assembly with automatically reconfiguration fixtures", IEEE Journal of Robotics and Automation, Vol. RA-1, No.2, June, 1985.

Table 1. A partial list of possible locator configurations

Locating edge combinations	locator configuration #1	locator configuration #2	locator configuration #3
three line segments	three locating towers (b)	three round locating pins (a)	two round locating pins (a) and one adjustable stop (c)
two line segments and one external arc	two round locating pins (a) and one half-Vee (d)	three round locating pins (a)	two round locating pins (a) and one adjustable stop (c)
one line segment and two external arcs	one round locating pin (a) and two half-Vees (d)	three round locating pins (a)	two round locating pins (a) and one adjustable stop (c)
one line segment and one external arc	one round locating pin (a) and one V-pad (e)	three round locating pins (a)	two round locating pins and one adjustable stop(c)
three external arcs	three half-Vees(d)	three round locating pins(a)	two round locating pins and one adjustable stop (c)
two line segments and one small internal circle	two round locating pins(a) and one diamond hole pin(g)		
two line segments and one large internal arc	three round locating pins(a)	two round locating pins(a) and one adjustable stop(c)	
one line segment and two small internal circles	one adjustable stop (c) and two diamond pins(g)		
two small internal circles	one round hole pin(f) and one diamond pin(g)		
three large internal arcs	three round locating pins(c)	two round locating pins(c) and one adjustable stop(c)	

Note: two line segments may degenerate into one; arc and circle may mean the same thing; and two half-Vees may be equivalent to one V-pad.

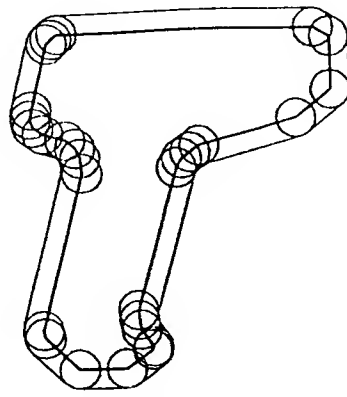


Figure 1. Expansion of workpiece boundary edges

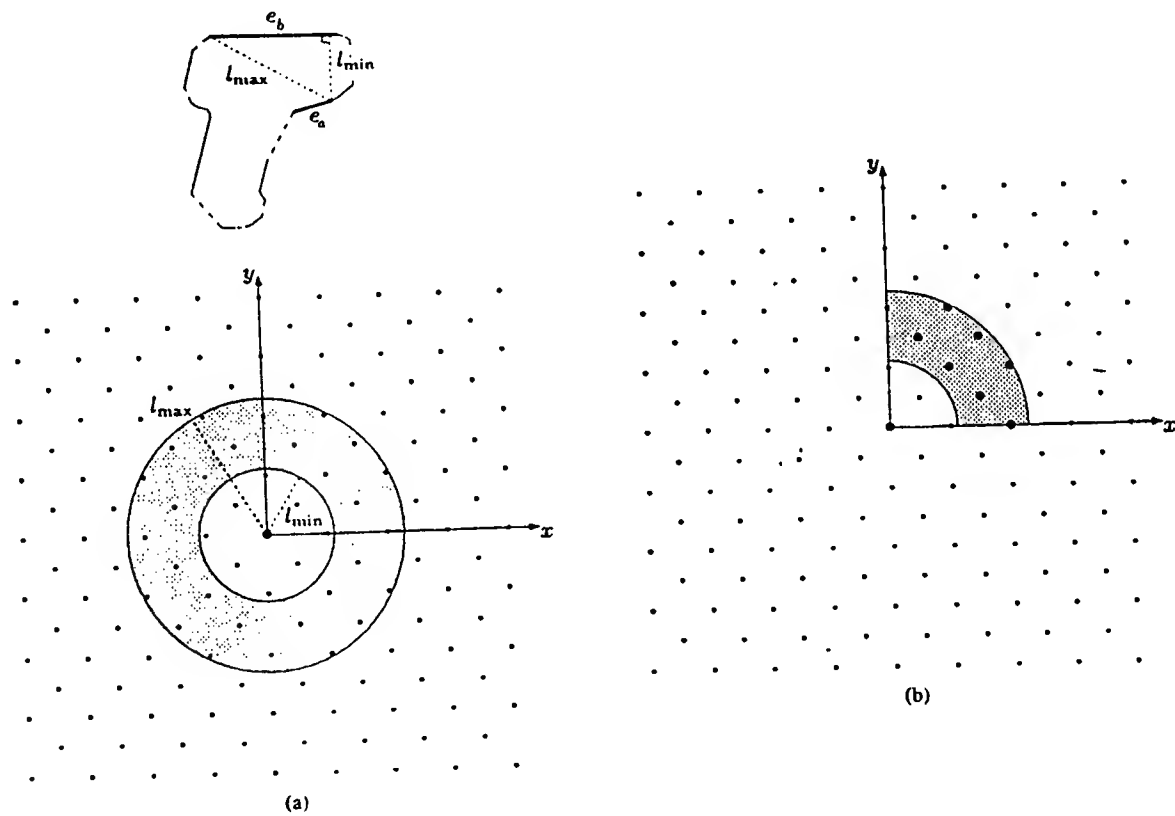


Figure 2. Search for locator positions to satisfy hole-alignment condition

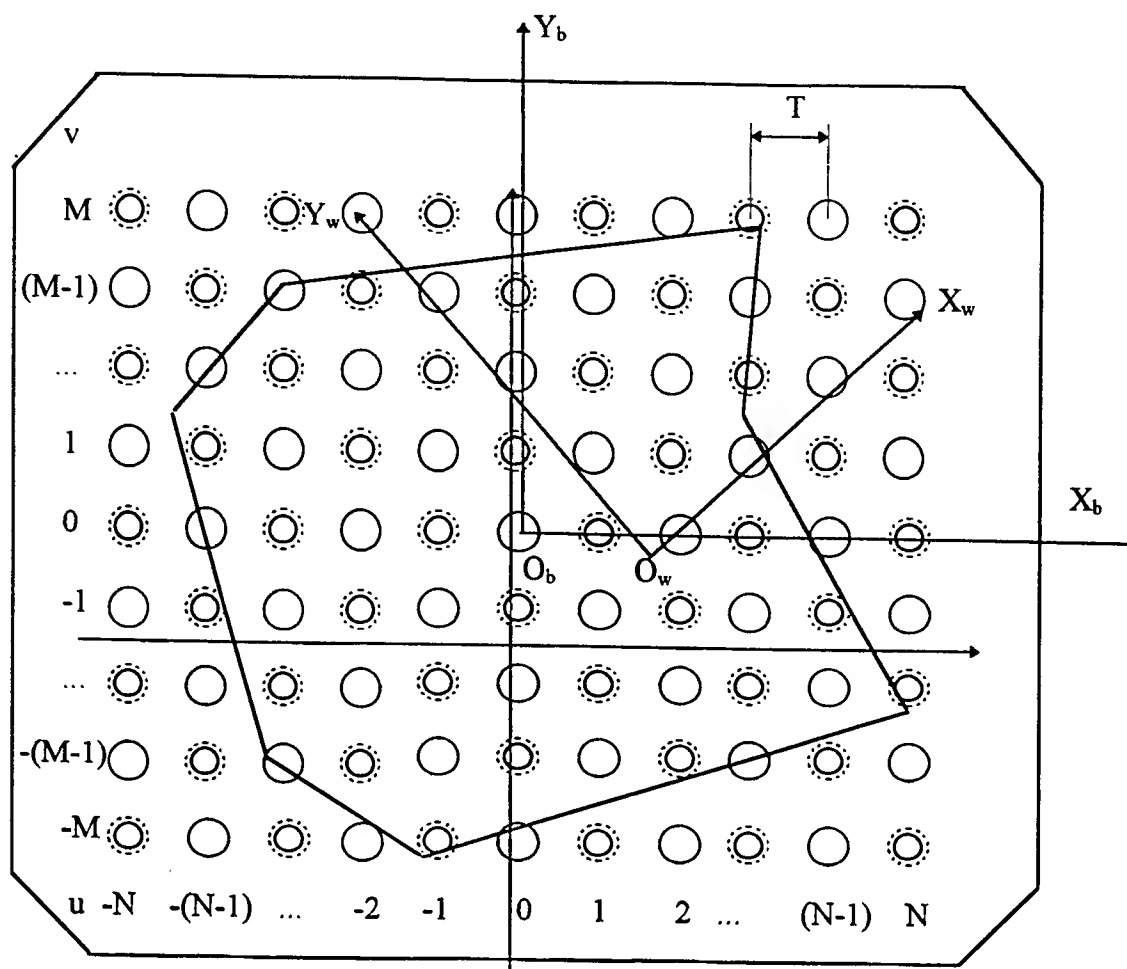


Figure 3. Base plate and workpiece coordinate systems

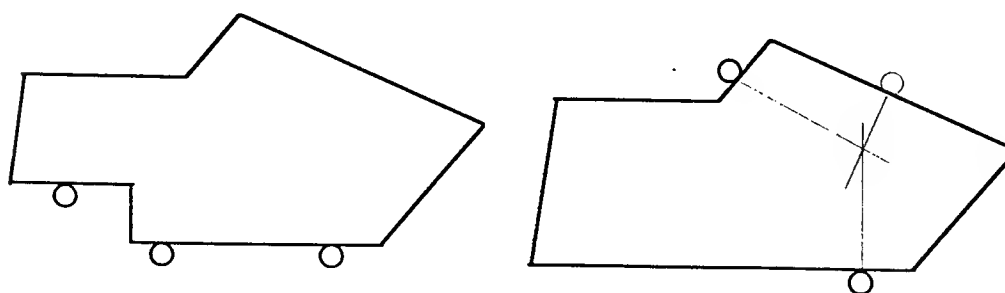
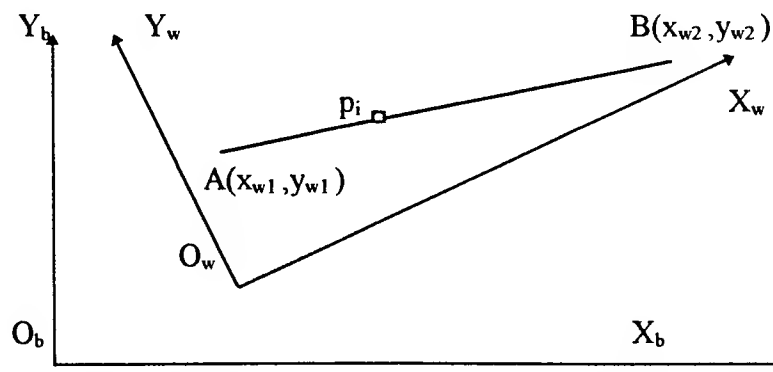
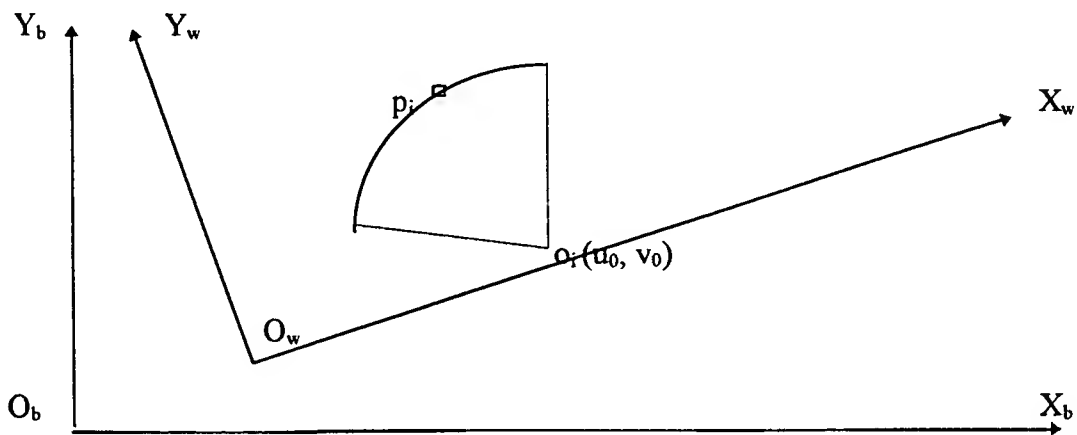


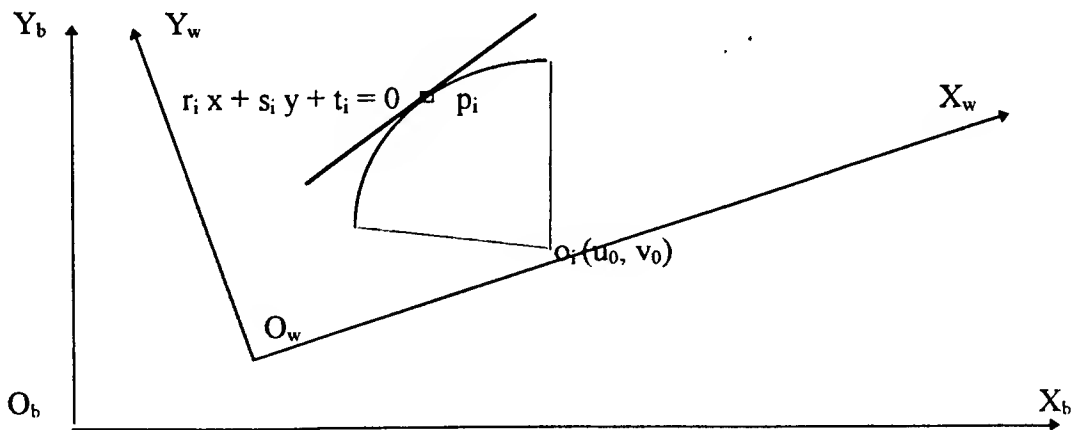
Figure 5. Examples of invalid locating designs



(a) A workpiece edge passing a locating point on base plate



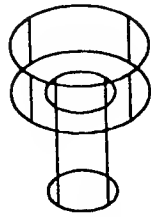
(b) A workpiece arc passing a locating point on base plate



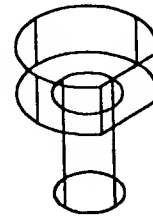
(c) A workpiece arc tangent to a locating line on base plate

Figure 4. Three types of assembly constraints

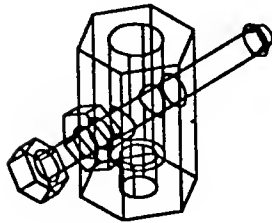




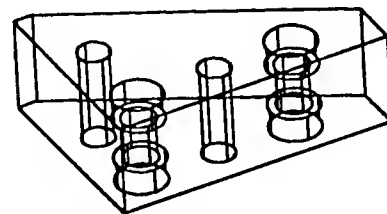
(a) round locating pin



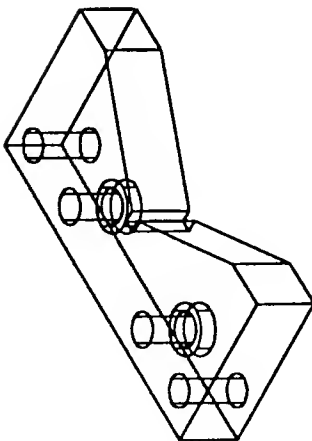
(b) locating tower



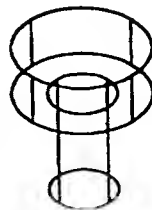
(c) adjustable stop



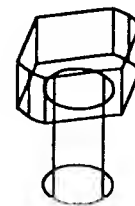
(d) half-Vee



(e) V-pad



(f) round hole pin



(g) diamond hole pin

Figure 6 Locators to be considered

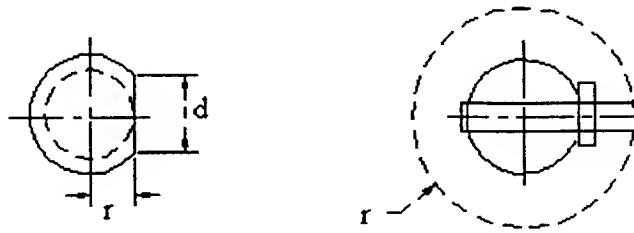


Figure 7. Simplified representation of locating tower (left) and adjustable stop.

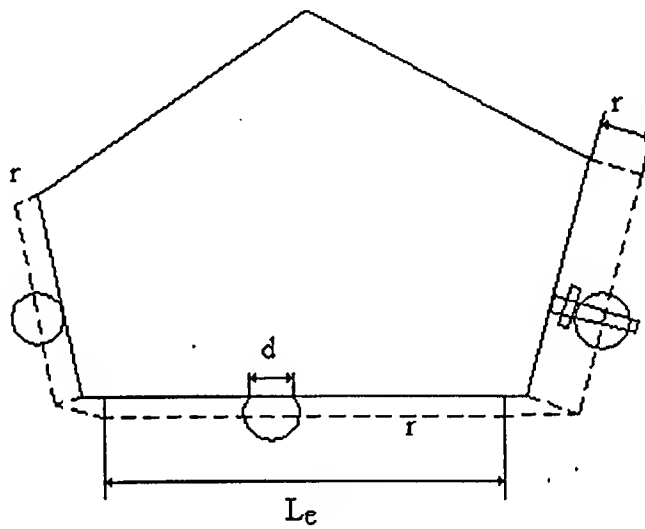


Figure 8. Application of different locating devices.

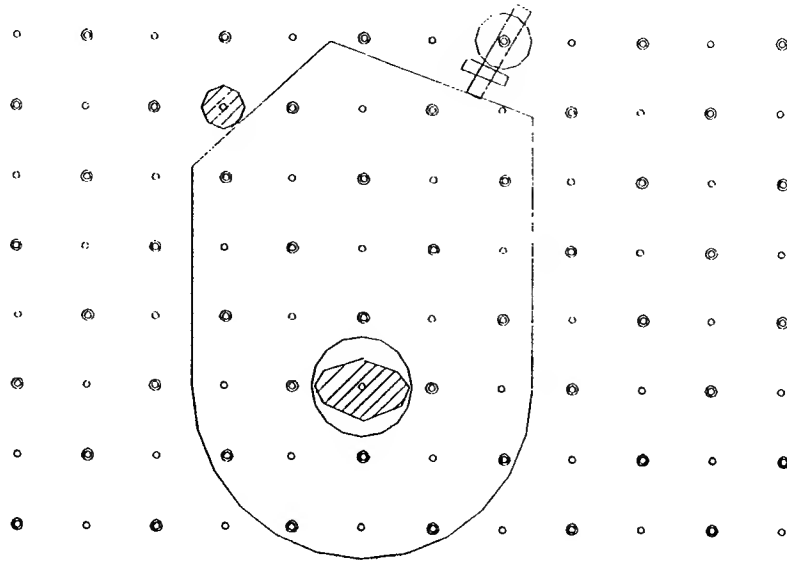


Figure 9.a Hole pin application.

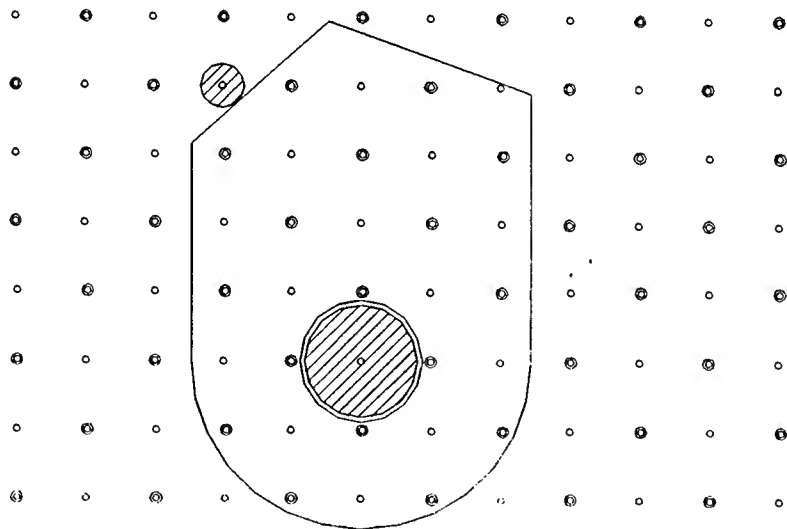


Figure 9.b Hole pin application

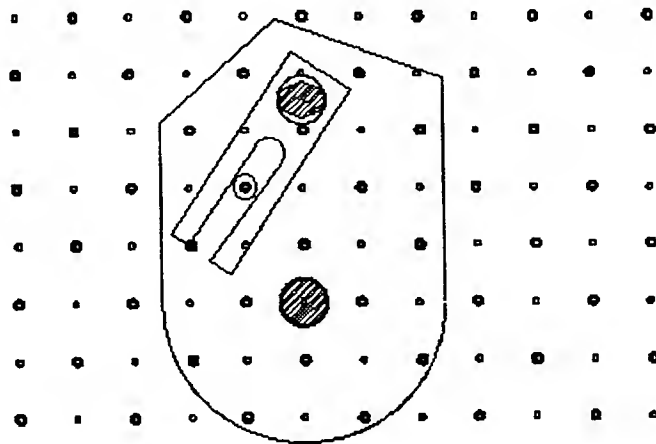


Figure 10. Pin-hole locating with adjustable bar

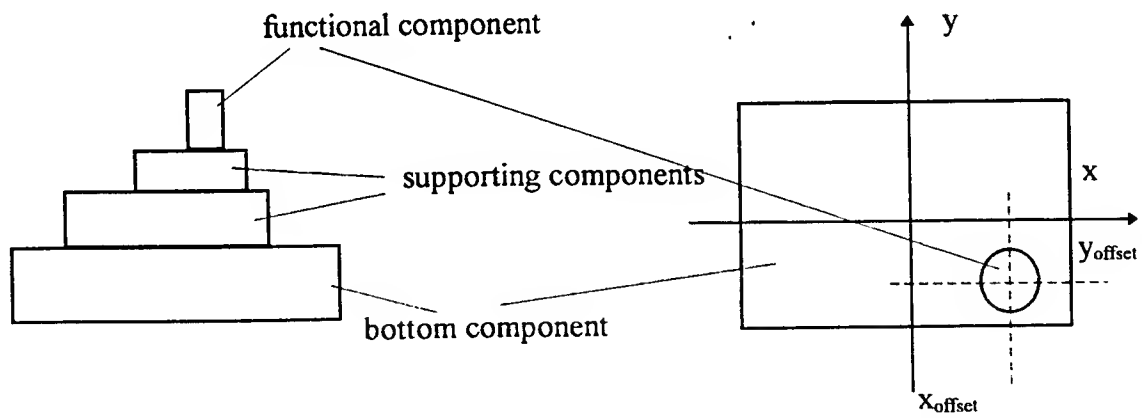


Figure 14. A sketch of fixture units

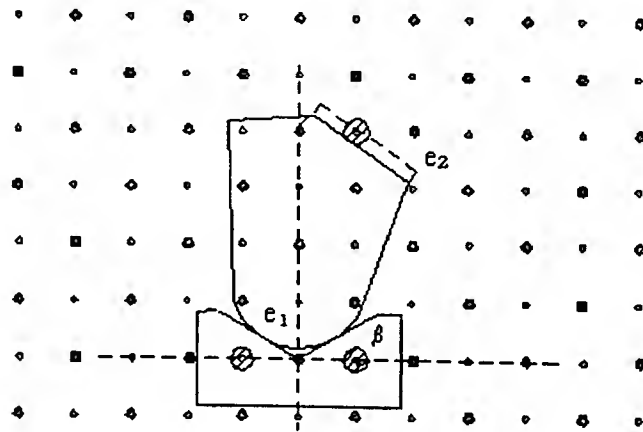


Figure 11.a V-block application.

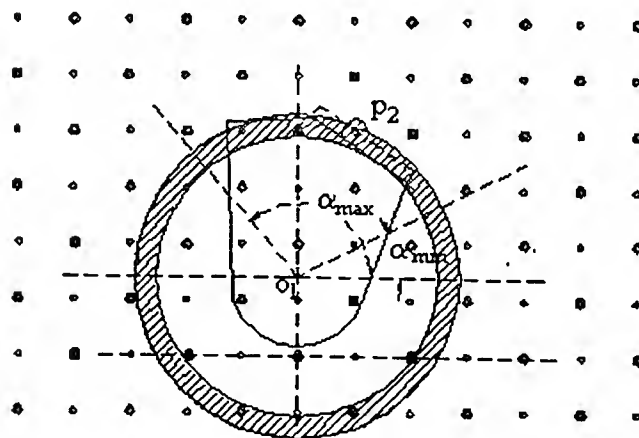


Figure 11.b V-block assembly analysis.

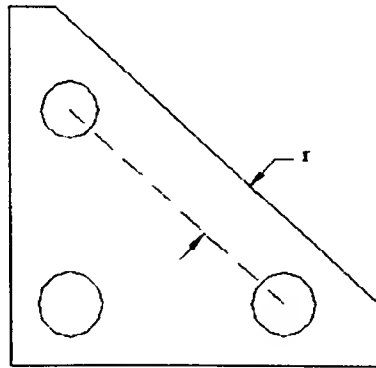


Figure 12.a Half-Vee

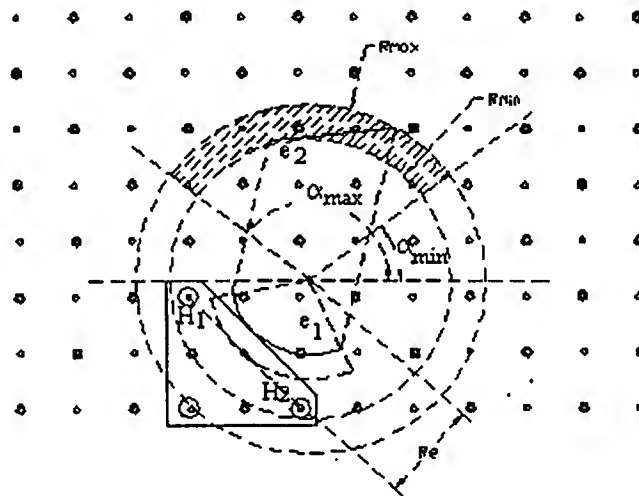


Figure 12.b Half-Vee assembly analysis (1)

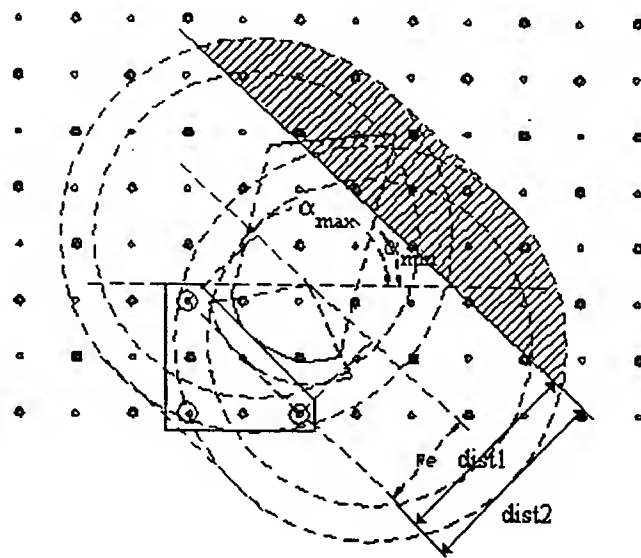


Figure 12.c Half-Vee assembly analysis (2)

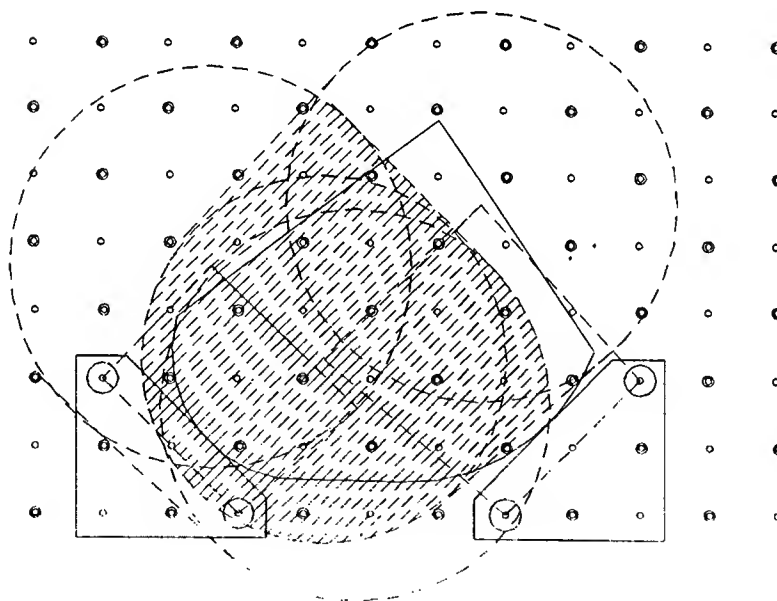


Figure 12.d Two half-Vee assembly analysis

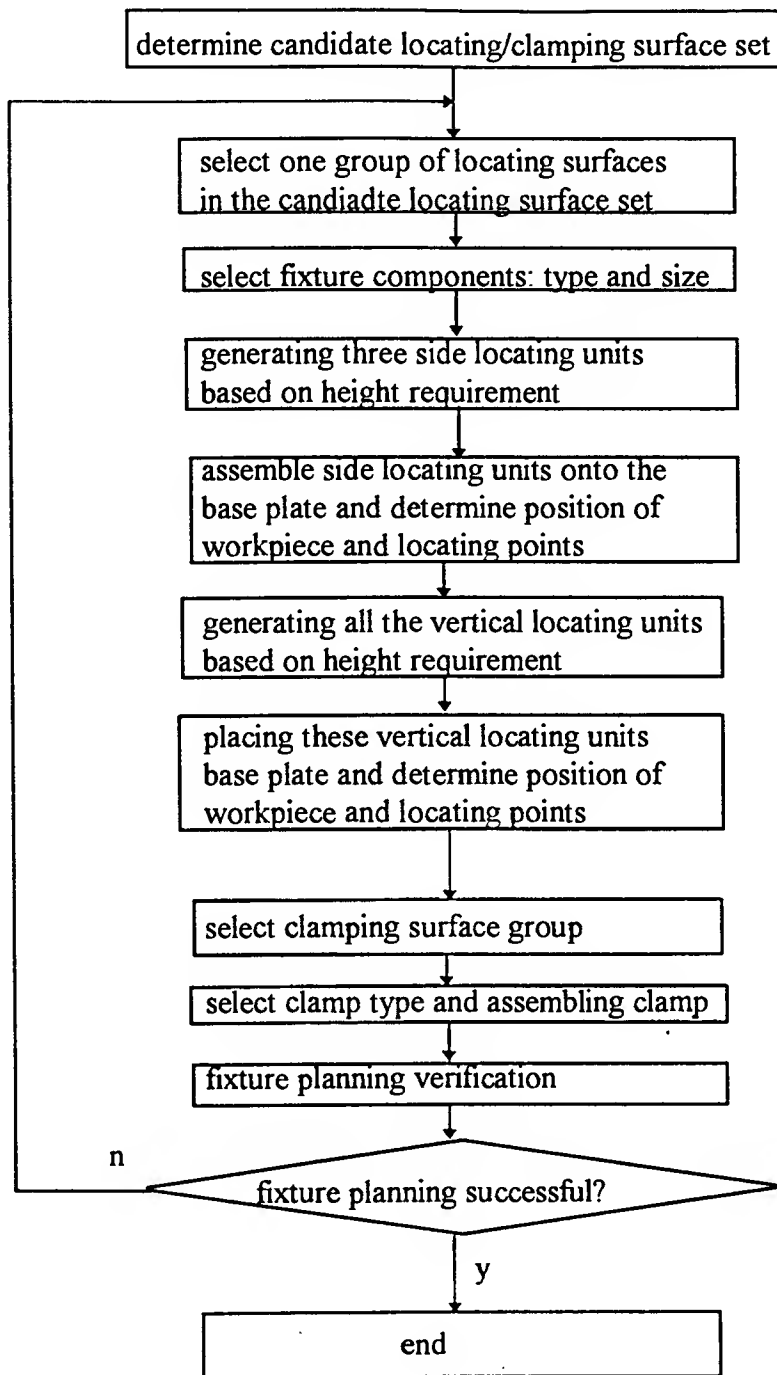


Figure 13. A diagram of 3-D automated modular fixture planning



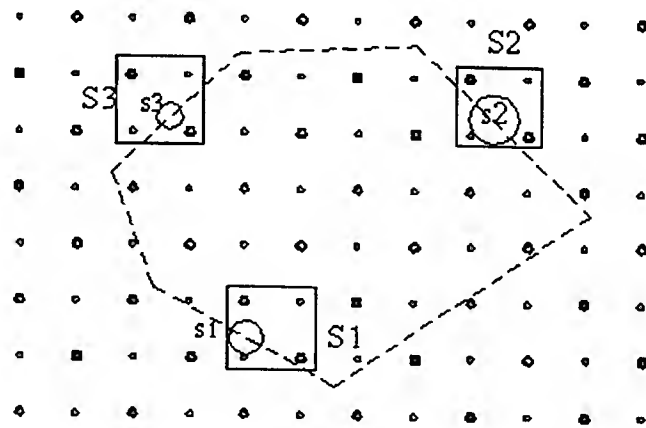


Figure 15. 3-D fixturing unit assembly

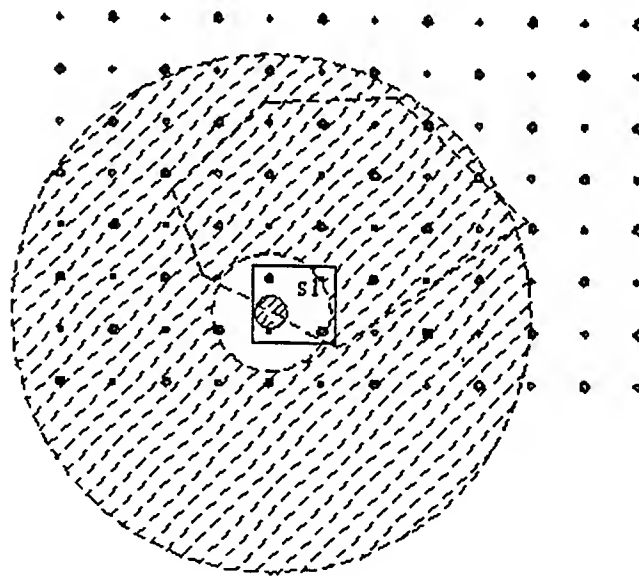


Figure 16. 3-D fixturing unit assembly analysis

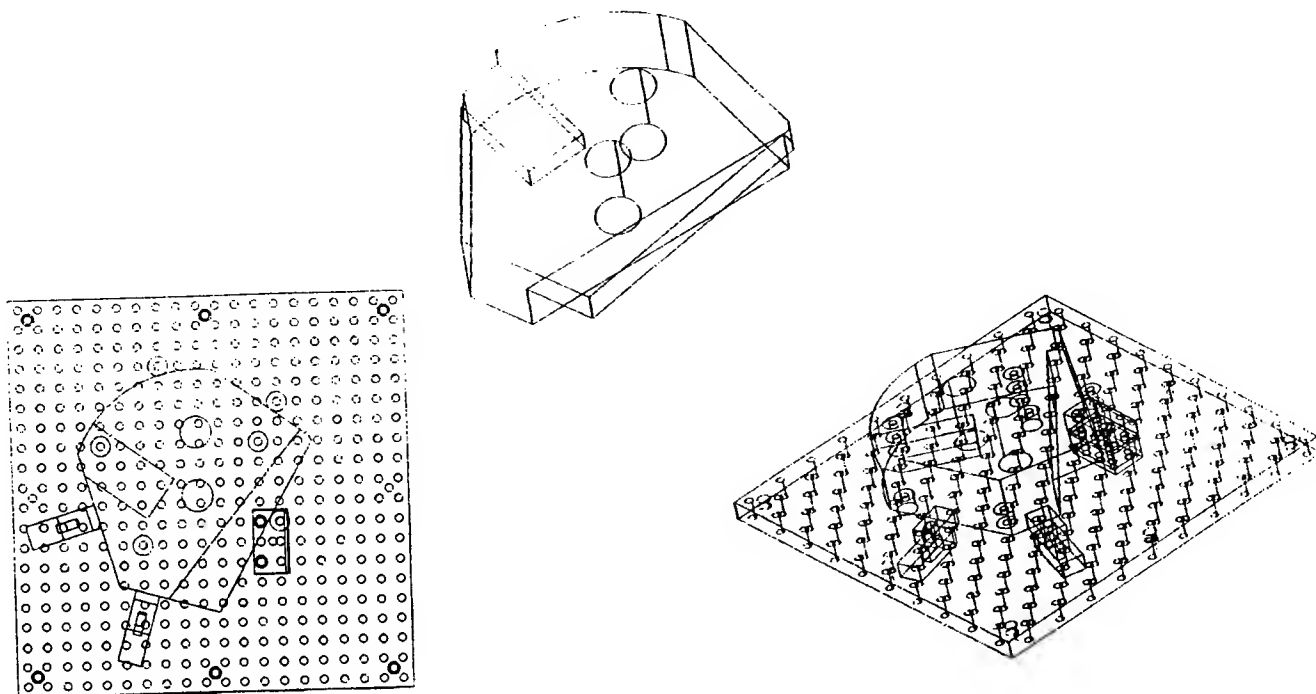


Figure 17.a Fixture design example #1

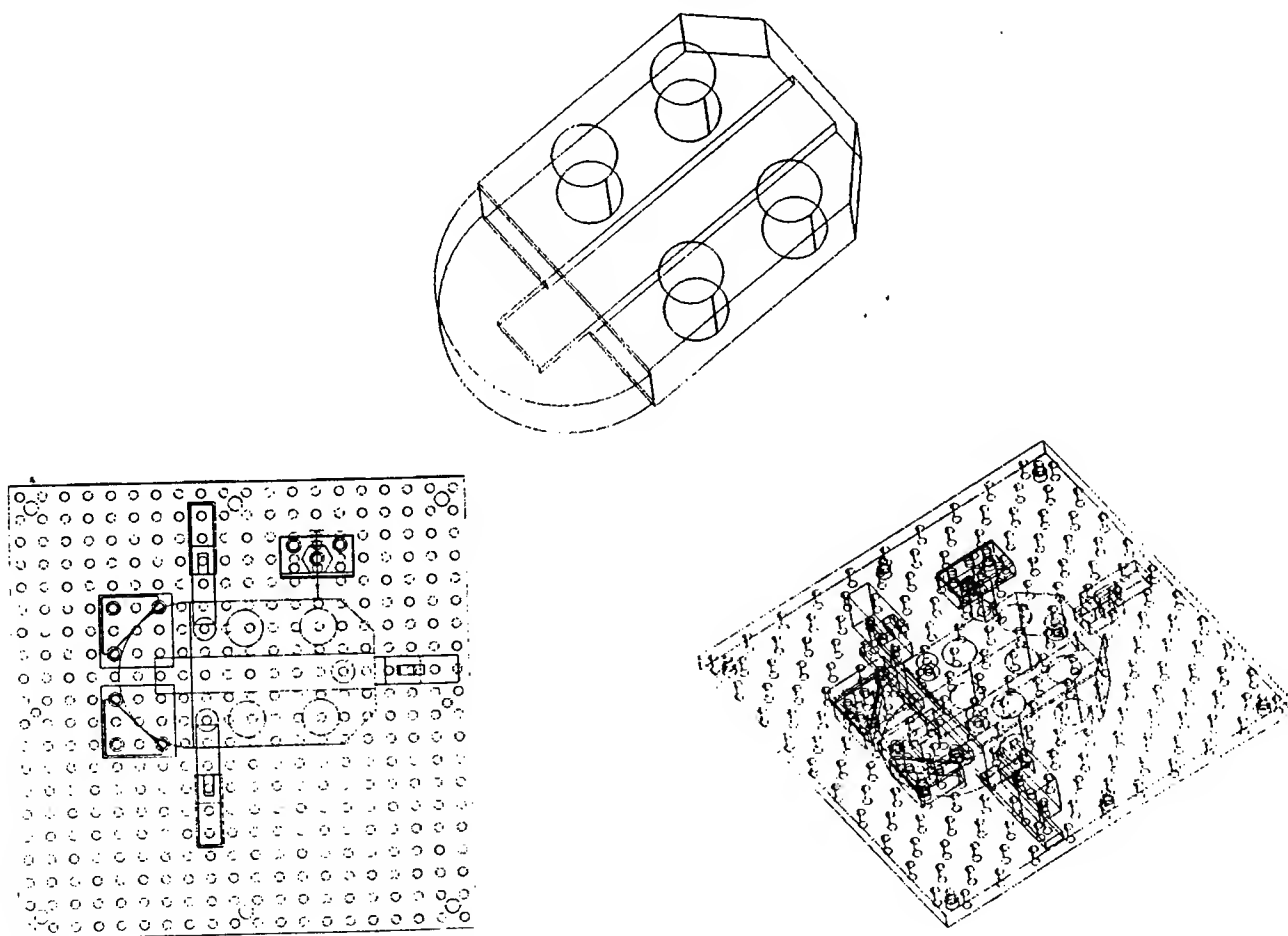


Figure 17.b Fixture design example #2

**John Schmalzel's report not available at time of publication**

**A DESIGN STRATEGY FOR PREVENTING HIGH CYCLE FATIGUE BY MINIMIZING SENSITIVITY  
OF BLADED DISKS TO MISTUNING**

**Joseph C. Slater  
Assistant Professor  
Department of Mechanical and Materials Engineering**

**Andrew J. Blair  
Graduate Research Assistant  
Department of Mechanical and Materials Engineering**

**Wright State University  
3640 Colonel Glenn Highway  
Dayton, OH 45435  
jslater@cs.wright.edu  
ablair@cs.wright.edu**

**Final Report for:  
Summer Faculty Research Program  
Wright Laboratory**

**Sponsored by:  
Air Force Office of Scientific Research  
Bolling Air Force Base, DC**

**and  
Wright Laboratory**

**December 1996**

---

A DESIGN STRATEGY FOR PREVENTING HIGH CYCLE FATIGUE BY MINIMIZING SENSITIVITY  
OF BLADED DISKS TO MISTUNING

Joseph C. Slater  
Assistant Professor  
Andrew J. Blair  
Graduate Research Assistant  
Department of Mechanical and Materials Engineering  
Wright State University

**Abstract**

Bladed disk assemblies in aircraft engines are prone to high cycle fatigue as a result of localization of vibrational energy. In order to prevent mode localization, design strategies for distributing dynamic stress over the entire system are examined. It is shown that stress reductions of up to 75% can be obtained via minor modifications of basic disk design.

# A DESIGN STRATEGY FOR PREVENTING HIGH CYCLE FATIGUE BY MINIMIZING SENSITIVITY OF BLADED DISKS TO MISTUNING

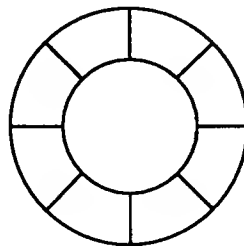
Joseph C. Slater  
Andrew J. Blair

## Introduction

The inside of a gas turbine is one of the harshest environments for mechanical systems in all of the man-made world. High temperatures, high forces and large dynamic loads make reliability design an exceptionally challenging task. Although jet engines statistically are very reliable overall, this reliability is likely to decline as aircraft are used more often to perform missions for which they were not designed. A design is only as robust as the qualification test it undergoes for approval. Many aircraft in the Air Force's aging fleet were designed for quite different roles than those they are currently being used for. This leads to the likelihood of performance degradation, reduced reliability, and shorter time to failure. A leading cause of failure in jet engines is fatigue, both low cycle and high cycle. Low cycle fatigue is a result of high loads being applied to an object over a relatively low number of cycles. On the other hand, high cycle fatigue is the result of lower loads being applied to an object for a large number of cycles. This is commonly the result of vibration over an extended period of time.

Mode localization in bladed disks is a vibration phenomenon where the symmetric mode shapes common to a perfectly symmetric (tuned), bladed disk degrade due to the introduction of slight variations within the blades. The presence of these slight variations is commonly referred to as mistuning. Modal motion occurs primarily in a few of the blades on the disk (often called the "rogue" blades). Since all of the modal energy is confined to a small number of blades, the amplitudes of the motion of these blades is greatly increased, resulting in greatly increased stresses, and reduced fatigue life.

In order to gain a thorough understanding of the localized behavior of a mistuned model, it is first necessary to understand the ideal mode shapes and frequency distributions of the tuned system. Take for instance the following bladed disk where the radial lines represent blades of the disk.

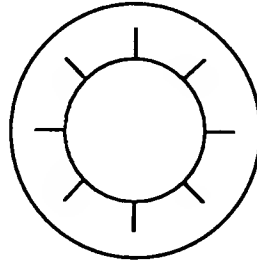


**Figure 1: Undeformed Bladed Disk**

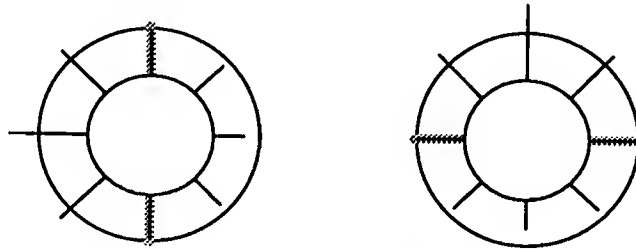
For the sake of illustration, shortening or lengthening of the radial line denotes deflection of a blade. The outer circle in the preceding figure represents the nominal position of the undeformed blade. In a real bladed disk, the

important deflections of the blades are usually bending, twisting, and combinations of the two – motions very similar to those of a cantilevered beam. For this illustrative example the deflection is represented as axial shortening or lengthening. This represents a simplification of the blade dynamics to those of a single beam-like mode.

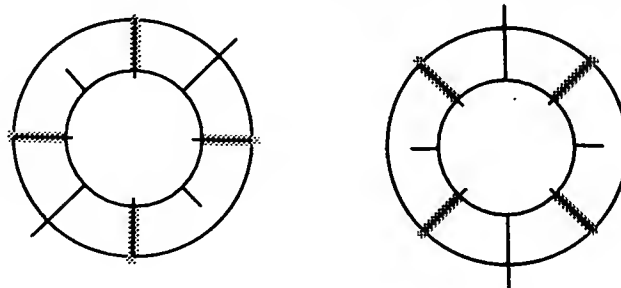
Since this is an eight degree of freedom model, eight linear modes can be expected. For the perfectly tuned system, these modes will be repeated and symmetric (except for two of the modes) as shown below. The gray lines represent the nodal diameters (lines) along which there is no deflection.



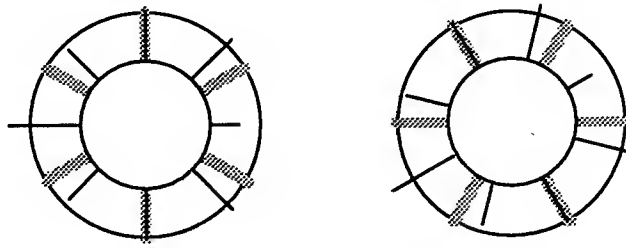
**Figure 2: First Mode (Zero nodal lines)**



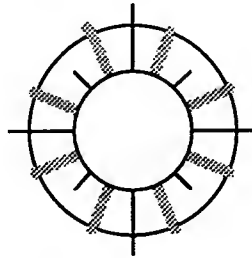
**Figure 3: Second and Third Modes (Repeated, with 1 nodal diameter)**



**Figure 4: Fourth and Fifth Mode (Repeated, with 2 nodal diameters)**



**Figure 5: Sixth and Seventh Mode (Repeated, with 3 nodal diameters)**



**Figure 6: Eighth Mode (Not repeated, 4 nodal diameters)**

It is reasonable to expect the modes to be repeated in this fashion when considering the symmetry involved in the structure. The preceding figures accurately depict the first of each type of modal group (beam bending and torsion). It is not difficult to extend this analogy of nodal diameters to the higher groups, though. The first group can be thought of as combinations of nodal diameters with zero nodal circles. As the modal group number increases, the number of nodal circles increases. So, the second bending mode would have one nodal circle, the third would have two nodal circles, and so on. In the present study, the first and second bending modal groups and the first torsional modal group were investigated.

The natural frequencies of each modal group in these systems are usually tightly grouped as well, especially in the lower groups. It is not uncommon for all of the frequencies in one group be within one percent of each other. Typically there is also a significant change in the mean frequency between modal groups.

The present study investigates the effects of mistuning on the mode shapes and frequencies of an eight bladed disk, as well as design alternatives to minimize these effects. Some previously used and accepted standards will be used to qualify the work. In addition, some new and unique measures are also formulated to summarize the findings.

### **Background**

Under certain conditions, due to the high flexibility of the blades relative to the disk, bladed disks can become very susceptible to mode localization as a result of blade mistuning (the variation of dynamic properties among the blades attached to the disk). The development of analytical methods for predicting the natural frequencies and mode shapes of tuned bladed disks has advanced sufficiently for use as design tools (although Swaminadham, Soni,



Stange, and Reed<sup>41</sup> demonstrate that significant work still needs to be done). What is lacking is the ability to predict the sensitivity of a design to mode localization that may result as a function of some specified mistuning, and knowledge of how to design bladed disks that are insensitive to blade mistuning.

Griffin and Hoosac<sup>17</sup> performed simulations of a simplified model of a bladed disk to generate a large number of simulation results from which to draw statistical conclusions. The model consisted of 72 three degree of freedom systems, representing the 72 blades of a bladed disk. The systems were coupled at the base by springs and each was connected to ground at the end mass by a dashpot to represent system damping. The end masses and their corresponding spring stiffnesses were varied to represent the effect of mistuning. The simulation demonstrated that, under the worst case scenario, the amplitude of a single blade could easily double the normal amplitude of a similarly excited tuned system. The shape of the scatter plot seems to indicate that the worst case scenario was not achieved and that the worst case would be catastrophic. Further analysis demonstrated that the range of blade natural frequencies and not the shape of the distribution of the blade natural frequencies was the dominant influence on blade amplitudes. It was also shown that the peak blade response occurred very near to the tuned natural frequency – validating the use of tuned analysis to design for natural frequency avoidance.

Griffin<sup>16</sup> uses a two degree of freedom model similar to that of Basu and Griffin<sup>2</sup>, but with more sophisticated coupling to represent aerodynamic coupling of the blades. The model was capable of predicting the scatter of blade amplitudes in some but not all cases. It was also found that systems that are sensitive to mistuning also tend to be numerically sensitive to modeling inaccuracies, suggesting that a deterministic approach to analyzing mistuning effects may not be possible with existing modeling techniques.

Valero and Bendiksen<sup>43</sup> developed a three degree of freedom blade model that incorporated rotation, shroud slippage, and friction. The shroud friction model assumes that the entire interface slips or sticks as a unit. This approach is then formulated as a linearized eigenvalue problem. The conclusion drawn was that the shroud interface angle alters the natural frequencies and the amount of friction damping observed. Surprisingly, the effects of mistuning were independent of the shroud interface angles. The mistuning effects tended to occur in the lowest modes where less deformation occurs in the hub (and thus the coupling between the blades is less apparent). It was also noted that the highly localized modes occurred when mistuning was highly concentrated in a few blades. In addition, the authors hypothesized that mode localization can be minimized by enhancing the interblade coupling through shrouds.

Two papers by Ewins<sup>12, 13</sup> represent standard reference material for understanding the modeling of bladed disks. Ewins<sup>13</sup> shows that under some mistuning conditions, blades may suffer stress levels as high as 20% greater than in a tuned system. However, rearranging the same blades on the same hub can minimize mode localization. An optimal arrangement of blade locations is proposed, although engineers<sup>22</sup> still have great difficulty in identifying the variations between a given set of blades for this purpose. Ewins<sup>13</sup> shows that many more resonant frequencies exist for mistuned bladed disk assemblies due to the splitting of repeated modes into independent modes. Ewins<sup>12</sup> shows that under some mistuning conditions, blades may suffer stress levels as high as 20% greater than in a tuned

system. These results are in contrast to those of Dye and Henry<sup>11</sup> and Whitehead<sup>45</sup>. Dye and Henry<sup>11</sup> show that almost a 3-fold increase in response amplitude can occur in the presence of mistuning, while Whitehead<sup>45</sup> analytically shows a theoretical increase of  $(0.5(1 + \sqrt{n}/2))$ .

Ewins and Han<sup>14</sup> used a two degree of freedom blade model to study the response of a 33-bladed disk. They show that, for this specific case, blade mistuning always results in an increase in blade amplitude and that the blade with the greatest mistune always suffers the greatest motion.

Yang and Griffin's<sup>46</sup> substructuring technique used the clamped-free modes of the blades to generate a reduced model of mistuned bladed disk assemblies. They showed that for a simple bladed disk assembly, the reduced model natural frequencies match the natural frequencies of the original finite element model almost perfectly, and, the peak forced response occurs at a frequency approximately 1% higher than the "true" frequency. For the mistuned case, the full and reduced models agree well. The exception was in a case where the tuned disk exhibits frequency veering, exactly the case in which mode localization occurs.

Irretier<sup>19</sup> applied a modified component mode synthesis to reduce a complete bladed disk finite element model to a smaller, more tractable problem. He showed that the manner of frequency shifting due to mistuning, and the corresponding change in mode shapes, were strongly dependent on the type of mistuning.

Kaza and Kielb<sup>20</sup> and Kielb and Kaza<sup>23</sup> used aerodynamically coupled single degree of freedom blade models for their bladed disk vibration analysis. They suggested that the effects of mistuning can be beneficial or adverse depending on the engine order of the forcing function. A significant result was that it may be possible to use designed mistuning to raise the blade flutter speed without seriously degrading the forced response, although the benefits of mistuning level off at about 5% mistune. Bendiksen<sup>3</sup> showed a similar result. Damping was shown to be much more effective when the blades are well tuned, which may cause problems when significant damping exists in the tuned system. A more sophisticated model showed many of the same results (Kaza and Kielb<sup>21</sup>).

Muszynska and Jones<sup>30</sup> developed a five degree of freedom blade model incorporating Coulomb shroud friction, Coulomb blade to hub friction and structural damping. Their model showed that mistuning increases the response amplitude and that appropriate design of friction dampers can reduce the response by as much as an order of magnitude as compared to a non-optimally designed friction damper. They also reported that the optimal damper design effectiveness is optimal for both the tuned and mistuned cases, although the amplitude for the mistuned cases were still higher than the amplitude for the tuned case. An unexpected effect was that the friction damping, due to its nonlinear nature, causes nonlinear coupling, inducing mode localization to some degree. Vakakis<sup>42</sup> has shown that when purely nonlinear coupling exists, mode localization can occur in the absence of mistuning.

Petrov<sup>35</sup> combined finite element, substructuring, transfer matrix, and dynamics compliance methods to develop a complex bladed disk model including shroud, joint, material damping, aerodynamic, and cable effects (for steam turbines). Isoparametric elements were used in the joint sections to model the complex geometries. A condensation technique was applied to reduce the size of the matrices<sup>36</sup>. The code shows that slight mistuning

drastically alters the clean transfer functions obtained for a tuned system, creating numerous resonances where only a handful previously existed.

Wei and Pierre<sup>44</sup> demonstrated that the sensitivity of a bladed disk to mode localization as a result of mistuning is directly related to the ratio of the mistuning strength to the coupling strength using a single degree of freedom blade model. They showed that the effects of mistuning are minimal when coupling is great. When coupling is weak, however, the bladed disk is very sensitive to mistuning. Thus, a bladed disk assembly that shows a great deal of motion of the hub when moving in a mode will be less susceptible to the effects of mistuning. Since more relative motion occurs in the hub in higher modes, it seems likely that higher modes should be less susceptible to blade mistuning.

Pierre and Murthy<sup>37</sup> and Pierre, Smith, and Murthy<sup>38</sup> included aerodynamic coupling of the blades in their perturbation approach to determination of the effects of blade mistuning. Since it was shown that the low coupling between the blades is the cause of the propensity for mode localization, Pierre and Murthy applied the approach of Wei and Pierre<sup>44</sup> in which the coupling treated as the perturbation of  $n$  originally independent blades with slightly different modal parameters. This is in contrast to procedures that include the mistuning as the perturbation of an originally tuned system. A heuristic explanation for why this may work is that when mode localization occurs, the blades act almost as independent structures. Since the nominal structure used by Wei and Pierre<sup>44</sup> was the set of independent blades, it is reasonable to expect that this should yield better results when localization occurs. Pierre and Murthy<sup>37</sup> also reported that blades similar in frequency tend to vibrate together in a localized mode, even when the blades between them do not show significant motion.

**A summary of these results leads to the following conclusions:**

- 1) Detailed finite element analysis of complex, tuned bladed assemblies is prone to large errors when mode localization occurs due to the intrinsic numerical ill-conditioning. However, degree of ill-conditioning is an indicator of the sensitivity of the design to mode localization.
- 2) Detailed finite element analysis (FEA) of complex, mistuned bladed assemblies can be extremely costly due to the inability to apply perfect symmetry relations. Accurate model reduction techniques are necessary.
- 3) Even if detailed FEA of complex, mistuned bladed disks could yield valid results, usefulness for design is questionable<sup>22</sup>.
- 4) The most promising method of gaining a detailed finite element model that is capable of incorporating the detailed effects of blade mistuning is to apply component mode synthesis in one of its various forms<sup>1, 6, 7, 8, 9, 15, 18, 24, 25, 32, 33, 40</sup>, as performed by Irretier<sup>19</sup>.
- 5) Mistuning has the greatest effect when coupling between the blades is weakest. Added coupling through shrouds is likely to reduce mode localization, minimizing the effects of bladed mistuning.
- 6) In addition, since higher sets of bladed disk modes tend to have greater coupling between the blades, mode localization may be a phenomenon of interest only with respect to the lower frequency sets of modes.

- 7) Dry friction can cause mode localization to occur in the absence of blade mistuning. Thus, in any assembly where friction dampers are used, the "effects" of blade mistuning exist regardless of how well the blade frequencies are tuned.
- 8) Blade mistuning is not guaranteed to cause significant mode localization. The same set of mistuned blades will exhibit symmetric modes or localized modes depending on the arrangement of the blades on the hub. Little is understood about why this is the case, or what way to order the blades to minimize this effect.
- 9) It is likely that blade mistuning can cause responses well above those reported in most studies. Monte Carlo simulations show distributions to have very sharp amplitude peaks indicating extremely large worst-case scenarios as opposed to soft peaks which would indicate milder worst-case scenarios.

### **Problem**

In bladed disk assemblies, the disk acts as a coupling device between the blades. As the stiffness of the disk increases, blade coupling decreases. It has been shown that weak interblade coupling leads to high levels of mode localization when blades are mistuned<sup>37,38,44</sup>. Mode localization also occurs in bladed disks as a result of their symmetry. Sets of axisymmetric modes combine to form a basis set from which drastically localized mode shapes can be generated. Bladed disk assemblies are traditionally designed to be symmetric for balancing. Two hypotheses are investigated in this work: 1) Decreasing the stiffness of the disk by varying the geometry and/or material composition will reduce mode localization due to mistuning, 2) Destroying the symmetry of the disk, yet maintaining balance, will reduce mode localization due to mistuning. Each of these was investigated separately, as well as combinations of the two.

A model of an eight bladed disk based on an experimental testbed in existence at Wright State University was constructed in ANSYS® using eight noded brick elements (Figures 10 and 11, p. 19). The model was designed to exhibit a propensity for mode localization similar to that in real bladed disk assemblies. The bladed disk model was adjusted to provide weak coupling between the blades—resulting in tightly packed sets of natural frequencies, eight modes in each. The blade deformation in the first set of modes is predominately a first beam bending mode, in the second set of modes it is predominately a first beam torsional mode, and in the third set of modes it is predominately a second beam bending mode.

<u>Frequency (Hz)</u>	<u>Deformation shape</u>	<u>Nodal lines, Nodal circles</u>
336.8	1 <sup>st</sup> Beam bending	0, 0
336.8	1 <sup>st</sup> Beam bending	1, 0
336.8	1 <sup>st</sup> Beam bending	1, 0
336.9	1 <sup>st</sup> Beam bending	2, 0
336.9	1 <sup>st</sup> Beam bending	2, 0
337.07	1 <sup>st</sup> Beam bending	3, 0
337.07	1 <sup>st</sup> Beam bending	3, 0
337.15	1 <sup>st</sup> Beam bending	4, 0
1411.4	1 <sup>st</sup> Beam torsion	0, 0
1411.4	1 <sup>st</sup> Beam torsion	1, 0
1411.4	1 <sup>st</sup> Beam torsion	1, 0
1411.5	1 <sup>st</sup> Beam torsion	2, 0
1411.5	1 <sup>st</sup> Beam torsion	2, 0
1411.8	1 <sup>st</sup> Beam torsion	3, 0
1411.8	1 <sup>st</sup> Beam torsion	3, 0
1412.0	1 <sup>st</sup> Beam torsion	4, 0
2066.1	2 <sup>nd</sup> Beam bending	0, 1
2066.7	2 <sup>nd</sup> Beam bending	1, 1
2066.7	2 <sup>nd</sup> Beam bending	1, 1
2072.8	2 <sup>nd</sup> Beam bending	2, 1
2072.8	2 <sup>nd</sup> Beam bending	2, 1
2081.4	2 <sup>nd</sup> Beam bending	3, 1
2081.4	2 <sup>nd</sup> Beam bending	3, 1
2084.6	2 <sup>nd</sup> Beam bending	4, 1

**Table 1: Modal characteristics of the model investigated.**

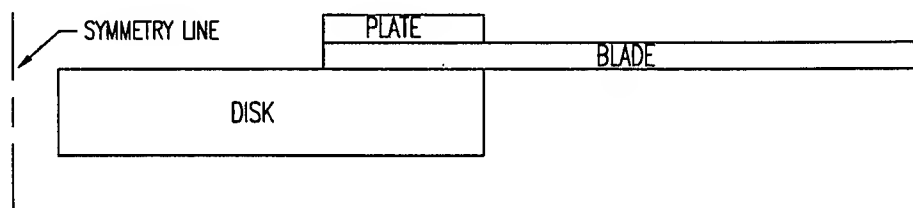
In all, seven different models were developed for this study. The aforementioned baseline model was symmetric with an axisymmetric disk stiffness, typical of traditional bladed disk design. In the second model, the symmetry of the disk was destroyed, yet balance was maintained. The bending stiffness of the interior portion of the disk was reduced in the third model. In the fourth model, the stiffness of the exterior of the disk was reduced while the interior portion of the disk remained unchanged. The fifth model was a combination of the second and third, where the disk was non-symmetric and the stiffness of the interior portion of the disk was reduced. The sixth model was a combination of the second and the fourth, where the disk was non-symmetric and the stiffness of the exterior portion of the disk was reduced. The seventh model was another variation of the non-symmetric disk. In this model, the disk was divided into eight equal sections. Each section was assigned a different stiffness (Young's Modulus), such that the disk would remain balanced.

Three types of mistuning were chosen for the investigation. The first two were; one percent of the mass of one blade added to the tip of one blade, and one percent blade mass removed from the tip of a single blade. This results in two cases of mistuning for the symmetric models. For non-symmetric models, however, each type (addition and subtraction) must be investigated on each of four different blades. This results in eight different cases of mistuning for the non-symmetric models. The preceding models representing, minor damage cases to an individual blade, are in agreement with accepted practice in prior studies<sup>37,38</sup>. The final type of mistuning cases is random

mistuning. Three random patterns were chosen such that the mass added to the tip of each blade could vary between plus or minus one percent of the mass of one blade.

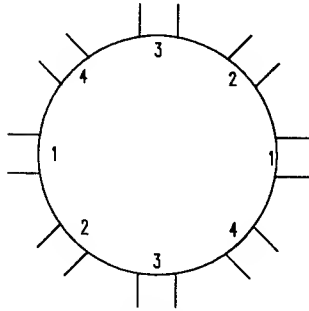
#### **Methodology- Models**

All models used in this study were variations of the symmetric, constant stiffness system. This standard system was constructed entirely of eight noded brick elements (ANSYS® element — Solid 45), having three degrees of freedom per node. This element was chosen over a plate element because of the overlapping condition of the disk, blades, and plates (Figure 7). The material properties used were that of mild steel: Young's Modulus ( $E$ ) = 200 GPa, density ( $\rho$ ) = 7800 kg/m<sup>3</sup>, and Poisson's Ratio ( $\nu$ ) = 0.3. The geometry for this model was created in AUTOCAD12®. Lines were also included in the geometry that allowed meshing to be performed in ANSYS®. A proper mesh in this study had two important properties. First, the mesh had to be symmetric. Due to numerical sensitivities, a non-symmetrical mesh could numerically induce undesirable mode localization characteristics that would not appear in a symmetrical, tuned disk. Second, the mesh had to consist entirely of "brick" elements — each element having eight independent nodes. These elements have a tendency to behave to "stiffly" when allowed to degenerate to "wedges" — elements in which all eight nodes are not independent.



**Figure 7: Mounting condition of each blade to the disk.**

The second model generated had a non-symmetric mass distribution and axisymmetric bending stiffness. A point mass element (ANSYS® element — Mass 21), was added to the edge of the disk at the centerline of each blade. The masses were added in a pattern that repeated after 180° in order to maintain the balance of the system (Figure 8). The procedure added ten percent of the mass of the disk to the system. This is an unacceptable increase in system mass. A more realistic implementation of this concept would decrease mass at some locations and increase mass in others, resulting in a smaller net change in mass.



1 = 0.5 % disk mass = 0.02825 Kg

2 = 1.0 % disk mass = 0.0565 Kg

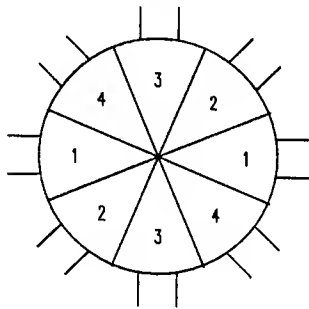
3 = 1.5 % disk mass = 0.08475 Kg

4 = 2.0 % disk mass = 0.113 Kg

**Figure 8: Mass distribution for the mass-added, non-symmetric model.**

The third and fourth models were symmetric, with reduced stiffness at the interior and exterior portions of the disk, respectively. For the simplicity of modeling, reducing the stiffness was accomplished by reducing Young's Modulus (E) to a value equal to half of its original value (100 GPa). In practice, similar reductions in stiffness would be accomplished by varying both design materials and geometry.

In conjunction with the second model, the third and fourth models were used to create models five and six. This gave two non-symmetric disks. One with reduced stiffness at the interior of the disk and one with reduced stiffness at the exterior of the disk.



1 = 100 % Young's Modulus = 200 GPa

2 = 80 % Young's Modulus = 160 GPa

3 = 60 % Young's Modulus = 140 GPa

4 = 40 % Young's Modulus = 80 GPa

**Figure 9: Stiffness distribution for the variable stiffness, non-symmetric model.**

When considering the second model, the investigators found it to be of considerable concern that, in practice, destroying the symmetry of the disk would inevitably lead to changes in the geometry. Not only would such a change mean the addition or subtraction of mass in certain locales, but it would also mean a considerable change in stiffness. It was this concern that brought about the seventh and final model. In this model, the disk was divided into eight equal sections. Four values of Young's Modulus were used in these sections in a repeated pattern similar to the mass distribution in the second model (Figure 9).

### **Methodology- Mistuning**

Three basic types of mistuning were chosen to best represent realistic challenges to the robustness of a bladed disk design. The first two were addition and subtraction of mass to a single blade. These were chosen to represent obstructions that may be “sucked into” a bladed disk system and either adhere to a blade or chip the blade in passing through the system. The final type of mistuning chosen was the random addition and subtraction of mass from each blade, representative of the small variance between blades due to manufacturing techniques and tolerances.

For the tuned systems, one percent of a single blade mass was added to the tip of each blade. To simulate the addition of one percent mass to one blade, the mass of the point element added to that blade was simply increased from one percent blade mass to two percent blade mass. To simulate the removal of mass from one blade, the one percent blade mass element was simply not added to that particular blade. For the symmetric models it was only necessary to investigate the results of adding mass to or removing mass from a single location (locations 1-4, **Figures 8 and 9**). However, for the non-symmetric models it was necessary to investigate the results of adding mass to or removing mass from four different locations (locations 1-4, **Figures 8 and 9**). Depending upon which location the mass was added to or removed from, the resulting mode shapes and frequency distributions of the system would be different.

Three random patterns of mistuning were also investigated. As mentioned before, the addition of one percent blade mass to each blade was chosen as the tuned, or nominal case. In determining the random patterns to be used, the mass to be added to any location was allowed to vary from zero to two percent of a single blade mass. This represents a tolerance of plus or minus one percent, with one percent of the mass of one blade being the nominal value. The random point mass distributions to be used were generated in MATLAB® using the randn function to generate three mistuning patterns, each with a normal distribution about the nominal value. The patterns generated follow in **Table 2**. This type of statistical representation (normal distribution about a nominal value) is widely used and accepted when incorporating manufacturing tolerances into design analysis.

<b>Blade Location</b>	<b>Random 1</b>	<b>Random 2</b>	<b>Random 3</b>
<b>#</b>	<b><u>E-3 Kg</u></b>	<b><u>E-3 Kg</u></b>	<b><u>E-3 Kg</u></b>
1	0.0168	0.3343	0.3413
2	0.0260	0.2867	0.4431
3	0.2579	0.4529	0.3710
4	0.3267	0.4119	0.1278
5	0.0037	0.2565	0.0231
6	0.1866	0.0448	0.3583
7	0.0325	0.3183	0.1598
8	0.2032	0.2025	0.3080

**Table 2: Random mistuning mass distributions.**



### **Methodology- Analysis**

Modal analysis was performed on all seven models for the tuned case and for each case of mistuning in ANSYS®, version 5.2. In each case, the inner surface of the disk was constrained to have zero displacement at all degrees of freedom. To minimize computer time, a lumped mass matrix formulation was used. The Subspace iteration method was used for eigenvalue extraction, with the tolerance for convergence checking set to 1E-5. The first twenty-four modes were extracted for all cases.

### **Methodology- Measures**

All cases of mistuning were investigated by application to each of the seven models. In each case the localized modes were compared to the nominal system. This resulted in the analysis of seven tuned systems and 59 mistuned systems. The results of the study were quantified using some standard measures. In addition, some new and unique measures were also developed to compare the tuned and mistuned cases.

Most studies concentrate on the first group of modes, corresponding to the first beam bending mode of a single blade. This is done for two reasons: first, and foremost, is the relative simplicity in mathematical formulation when assuming one general shape of deformation; second, is that mode localization tends to occur to a greater degree in the lower modes<sup>44</sup>. When looking at only the first beam bending modes, it is sufficient to quantify the results by looking at the displacement at the tips of the blades. In this particular mode of deformation there is a direct correlation between the tip displacement of a blade and the amount of stress at its root. This is not true for higher modes. To properly investigate the higher modes, some measure of energy or stress was needed.

ANSYS® 5.2 has the built in capability to determine relative Principal and Von Mises Stresses at each node. It is important to notice that these stresses are qualified as relative because there is no forcing of the system in modal analysis. This capability was utilized in determining a meaningful measure for presenting the results of this study. It is intuitive that the blade with the highest value of Von Mises Stress (VM) will be the one where the greatest mode localization takes place. To obtain a true measure of mode localization, this maximum value of stress was compared to the average value of stress for the entire system, yielding a stress ratio  $R_t$ . Here,  $R_t$  represents the stress ratio for the tuned case and  $R_m$  represents the stress ratio for the mistuned cases.

$$(a) \quad R_t = \frac{VM_{\max, \text{tuned}}}{VM_{\text{avg}, \text{tuned}}}$$

$$(b) \quad R_m = \frac{VM_{\max, \text{mistuned}}}{VM_{\text{avg}, \text{mistuned}}}$$

Finally, and most importantly, the mistuned response ( $R_m$ ) of each model was compared to that of the other models for each type of mistuning. This is done graphically and can be found in Figures 13-24 (pp. 20-31). It is from these graphs, that the conclusions are drawn by the investigators.

Some mathematical manipulation was required to justify the comparison of results between different models. ANSYS® normalizes the resulting eigenvectors (mode shape vectors) by the mass matrix in the following manner:

$$\{\Phi\}_i^T [M] \{\Phi\}_i = 1$$

where,  $\{\Phi\}_i$  is the  $i$ th mode shape eigenvector, and  $[M]$  is the system mass matrix. The relative stresses are then calculated using these normalized eigenvectors. This creates a problem in comparing eigenvectors or stress vectors of different models, since for different models, these vectors are normalized to different mass matrices. To remedy this situation, both the eigenvectors and the stress vectors were normalized to unit length.

When mistuning occurs in a given system, it is important to realize that groups of modes remain grouped. In fact, the mistuned mode shapes are shown to be linear combinations of the tuned mode shapes. This is verified by the following calculation.

$$\begin{matrix} [\Phi_t]^T [\Phi_m] = [\Psi] \\ 24 \times n \quad n \times 24 \quad 24 \times 24 \end{matrix}$$

where,  $[\Phi_t]$  and  $[\Phi_m]$  are the matrix of tuned eigenvectors (mode shapes) and matrix of mistuned eigenvectors (mode shapes), respectively. The variable  $[\Psi]$  is defined as the transformation matrix and  $n$  is the number of degrees of freedom. This results in an  $m \times m$  matrix, where  $m$  is the number of modes examined ( $m=24$ , in this case). The transformation matrix can then be divided into three submatrices, corresponding to the three modal spaces (modal groups) being investigated.

$$[\Psi]_{24 \times 24} = \begin{bmatrix} \Psi_1 & & \\ & \Psi_2 & \\ & & \Psi_3 \end{bmatrix}$$

$\begin{matrix} \Psi_1 & \Psi_2 & \Psi_3 \\ 8 \times 8 & 8 \times 8 & 8 \times 8 \end{matrix}$

The determinate of each of the three submatrices should have a magnitude of one, if the mistuned mode shapes are indeed linear combinations of the corresponding tuned mode shapes for each modal group. Each submatrix can also be examined for the contribution of each of the tuned mode shapes to each of the mistuned mode shapes. Here, each column represents a mistuned mode shape, and each row represents a tuned mode shape. In the present example ( $\Psi_1$ , model 1, mistuning case 1), it is easily seen that the second mistuned mode shape is comprised mainly of the first and third tuned mode shapes. The seventh mistuned mode shape is comprised almost entirely of the sixth tuned mode shape, and so on.

$$\Psi_1 = \begin{bmatrix} 0.315 & 0.734 & 0.451 & 0.360 & \sim 0 & 0.159 & 0.005 & -0.040 \\ 0.167 & 0.371 & -0.892 & 0.173 & \sim 0 & 0.089 & -0.003 & -0.025 \\ 0.510 & -0.570 & \sim 0 & 0.582 & \sim 0 & 0.268 & \sim 0 & -0.070 \\ 0.505 & \sim 0 & 0.015 & -0.689 & \sim 0 & 0.509 & -0.007 & -0.103 \\ \sim 0 & \sim 0 & \sim 0 & \sim 0 & -1 & \sim 0 & \sim 0 & \sim 0 \\ -0.007 & 0.003 & 0.005 & 0.007 & \sim 0 & \sim 0 & -1 & -0.010 \\ 0.490 & -0.003 & \sim 0 & -0.135 & \sim 0 & -0.754 & \sim 0 & -0.417 \\ 0.343 & -0.003 & -0.003 & -0.075 & \sim 0 & -0.261 & -0.012 & 0.900 \end{bmatrix}$$

Along with the resulting change in mode shapes, it is equally important to examine the resulting changes in the natural frequencies of the mistuned system relative to those of the tuned system. In the tuned system, especially systems in which the disk is symmetric, all natural frequencies of each group are within a very small percentage of each other. In the mistuned models, some of the natural frequencies in each group can move considerably away from the mean frequency of that group (this phenomenon is commonly known as eigenvalue veering).

Mode #	Tuned Hz	Cases 1-4 Hz	Cases 5-8 Hz	Case 9 Hz	Case 10 Hz	Case 11 Hz
1	336.8	330.64	336.8	334.74	331.49	331.74
2	336.8	336.8	336.8	336.56	332.55	333.6
3	336.8	336.8	336.85	338.01	334.54	333.92
4	336.9	336.85	336.9	338.45	334.96	334.36
5	336.9	336.9	337	342.68	335.8	335.24
6	337.07	337	337.07	342.83	336.59	339.18
7	337.07	337.07	337.13	343.15	338.03	340.05
8	337.15	337.13	343.59	343.49	342.34	342.94
9	1411.4	1411.3	1411.4	1411.4	1411.3	1411.3
10	1411.4	1411.4	1411.4	1411.4	1411.3	1411.4
11	1411.4	1411.4	1411.4	1411.4	1411.4	1411.4
12	1411.5	1411.5	1411.5	1411.5	1411.5	1411.5
13	1411.5	1411.5	1411.5	1411.6	1411.5	1411.5
14	1411.8	1411.8	1411.8	1411.9	1411.8	1411.8
15	1411.8	1411.8	1411.9	1411.9	1411.8	1411.9
16	1412	1412	1412	1412	1412	1412
17	2066.1	2038.4	2066.5	2059.6	2040.9	2042.6
18	2066.6	2066.5	2066.8	2073.1	2051	2055.1
19	2066.7	2066.8	2069.5	2078.6	2057.8	2056.1
20	2072.6	2070.3	2072.8	2081	2061.3	2059
21	2072.8	2072.8	2076.9	2104.3	2070.6	2067.2
22	2081.4	2078.5	2081.6	2108.9	2072.5	2088.6
23	2081.6	2081.6	2083.7	2113.9	2082.8	2091.3
24	2084.6	2083.9	2114.1	2115.4	2106.7	2110.9

**Table 3: Frequency table for Model 1 (symmetric mass distribution, axisymmetric stiffness).**

In the case of the tuned symmetric disk, the frequencies in each group are all within one percent of each other. This is no longer the case when the system is mistuned. In mistuning cases 1-4 (addition of mass to one blade), the first frequency of the first mistuned group deviates considerably from the rest of its group. In cases 5-8 (removal of mass from one blade), it is the last frequency of that group that deviates. There is no single frequency "leaving" the group in the random mistuning cases (9-11). Instead, the frequencies of that group are dispersed over a larger range. It should be noted here that, in the case of mass addition or removal, the mode in which frequency deviates most from the group is also the mode which exhibits the strongest mode localization. A complete set of frequency tables for each model can be found in **Tables 4-10** (pp. 32-37).

## **Results**

It was the goal of the investigators to minimize the detrimental effects of mode localization by altering the symmetry and/or stiffness of the disk. Individual blade damage (addition and subtraction of mass) were

investigated, as well as random mistuning representative of small manufacturing variances. The results of this study are presented in **Tables 4-10** (pp. 32-37) and most significantly, **Figures 12-23** (pp. 20-31). The results of this study may prove to be very useful in the future design of bladed disk assemblies, especially when the higher modes of vibration are of particular concern.

In the examination of the stress ratio, it is important to first look at the tuned models. In the case of the symmetric disk with axisymmetric stiffness (the baseline model to which all design modifications are compared), stress ratio values ( $R_t$ ) range between: 9-11 for the first modal group, 3-7 for the second modal group, and 5-8 for the third modal group. The proposed design changes offered little improvement in the stress ratio for the tuned case. In fact, the selected methods of destroying the disk's symmetry actually induced mode localization, increasing the value of the stress ratio. However, this result is not of dire consequence to future design, because the perfectly tuned case will never be a practicality.

The results of the mistuned cases are best summarized by looking at each case, one modal group at a time. The first modal group corresponds to the first beam bending shape of an individual blade. In the cases of blade damage, one mode of this group showed a significant increase in stress ratio. When mass was added, the first mode of the group showed a significant increase in stress ratio. However, when mass was removed, the last mode of the group showed a significant increase in stress ratio. This mode localization is clearly visible when the same mode is plotted for the tuned and mistuned cases, using the same deformation scale for each (**Figures 10 and 11**, p. 19). None of the proposed changes offered a significant improvement in the response of this highly localized mode. The techniques employed to destroy the symmetry of the disk even caused one or more other modes to show an increase in stress localization. For the three random cases of mistuning, all eight modes of the group were found to have high levels of stress localization. Destroying the symmetry of the disk causes sporadic improvement of stress levels throughout this group. Most importantly, reducing the stiffness of the inner portion of the disk, results in consistent lowering of the stress level throughout the entire group (relative to the mistuned baseline system).

The second modal group corresponds to the first torsional group of an individual blade. For all cases of mistuning, the stress levels of this group were only slightly raised. Again the destruction of symmetry was shown to have a negative affect on the system. Although some sporadic improvement was shown by changing both the interior and exterior stiffnesses of the disk, no consistent improvement was shown in this modal group by implementing any of the proposed design changes.

The most interesting results are found in the examination of the third modal group, corresponding to the second beam bending mode of an individual blade. In the cases of blade damage, there was only a single highly localized mode, similar to the results obtained for the first modal group. The only proposed design change that did not offer significant improvement in the blade damage cases was the non-symmetric (mass added), axisymmetric disk stiffness model. Reducing the interior disk stiffness seemed to trigger the most improvement. When this was

coupled with the destruction of symmetry, via the addition of mass, the maximum stress ratio dropped to as little as one-fourth that of the mistuned baseline system. In the random mistuning cases all proposed design changes offered some improvement, although the results of the mass added destruction of symmetry were rather inconsistent. Again, the reduction of the interior disk symmetry seemed to be the most beneficial design modification. Implementing both the reduction of the interior disk stiffness and mass added destruction of symmetry yielded the best results. Here, the stress ratio dropped to as little as one-half that of the baseline system.

### **Conclusion**

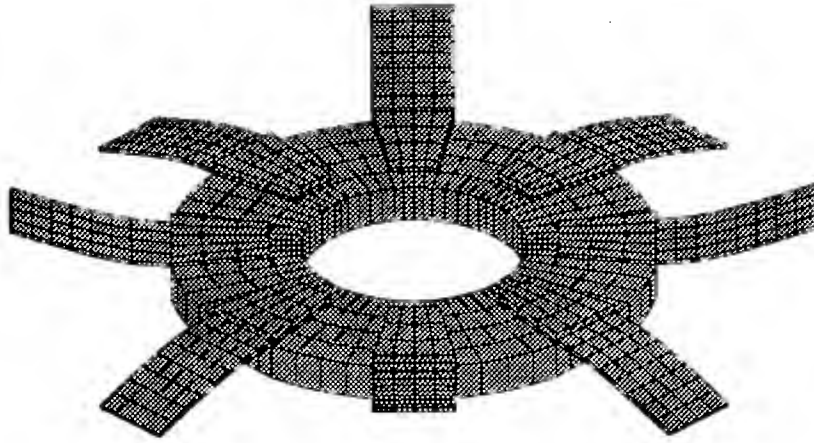
Reducing the interior disk stiffness relative to the exterior disk stiffness can dramatically improve the performance of bladed disk systems in the presence of mistuning. Upon examination of the different responses in the third modal group, one can also conclude that there is some promise in the prospect of destroying the symmetry of the disk. Although no great improvement was shown in the responses of the first and second modal groups, the dramatic improvement shown in the third modal group is enough to warrant consideration for the design changes proposed.

Only one pattern of mass addition was used in this study to represent a non-symmetric disk. The addition of this mass, in itself, was shown to induce significant levels of mode localization. The most promising results were shown when destroying the disk's symmetry, via the addition of mass, was coupled with a reduced interior disk stiffness. It is quite possible that the amount of mass added to create a non-symmetric disk was too drastic. The investigators speculate that the positive effects of reducing the interior disk stiffness was enough to overcome the negative effects caused by destroying the symmetry. Had the amount of mass added to the disk been less severe, the implementation of a non-symmetric disk, in itself, may have proven beneficial.

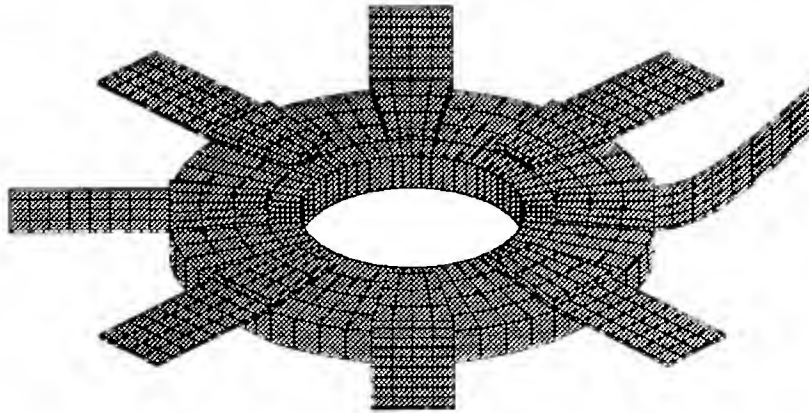
### **Future Study**

The present study featured seven different types of models. Two types, disks with reduced interior stiffness and mass added non-symmetric disks, yielded promising results and are worthy of further investigation. Based on the conclusions of previous authors, it is the authors' belief that reducing the stiffness of the outer portion of the disk is also worthy of further investigation. Future work should include:

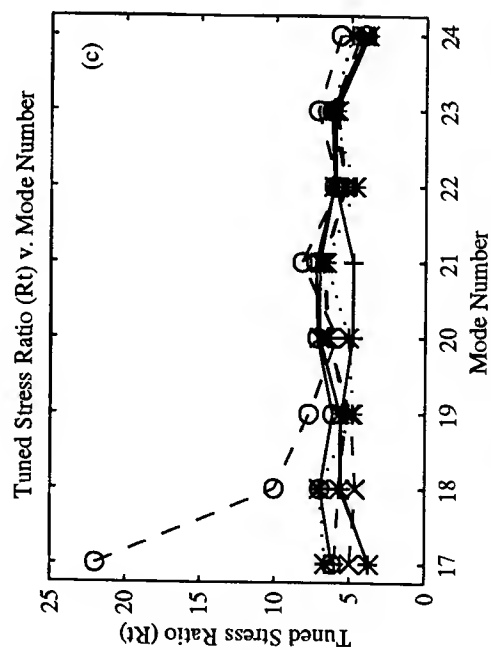
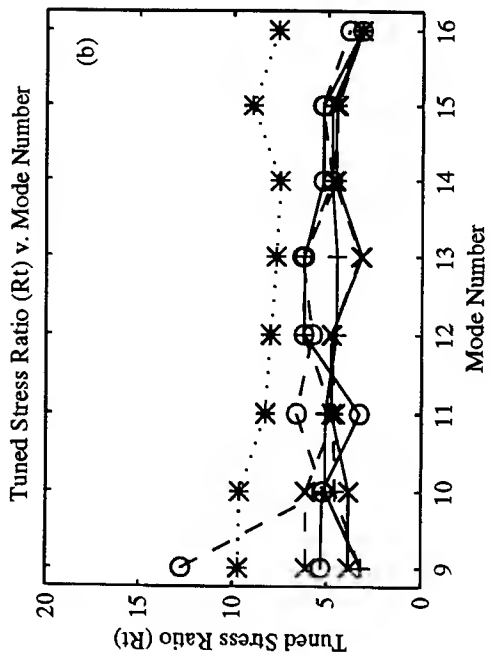
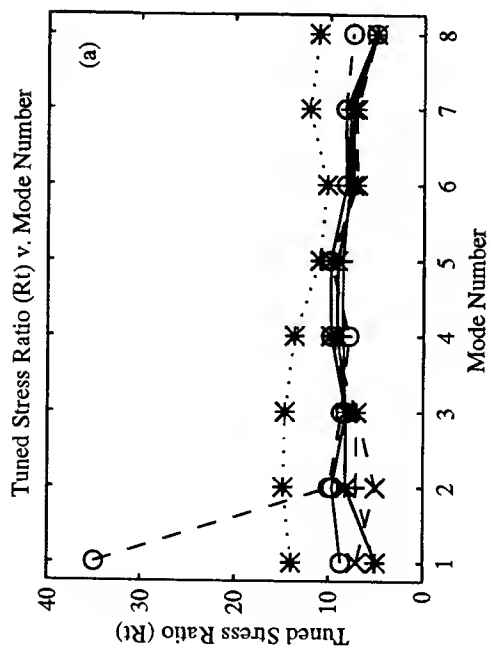
1. Investigation of different levels of relative stiffness between the interior and exterior portions of the disk.
2. Investigation of different patterns and mass amounts used to destroy the symmetry of the disk.
3. Investigation of combinations of the previous two.
4. Formulation of new quantitative measures to summarize investigative results.



**Figure 10: Model 1 (baseline), eighth mode, tuned case**

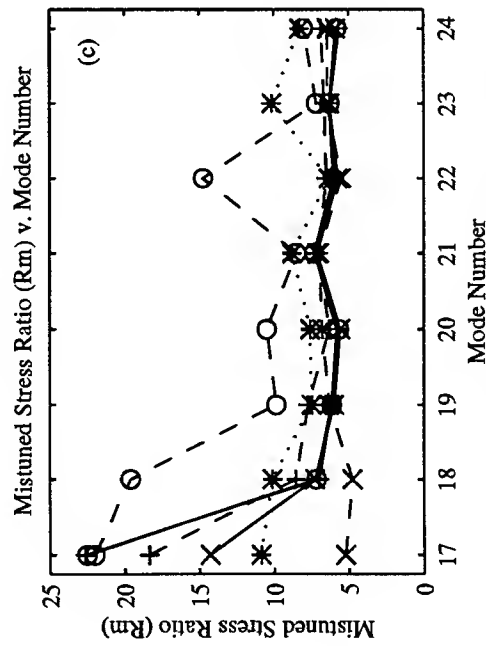
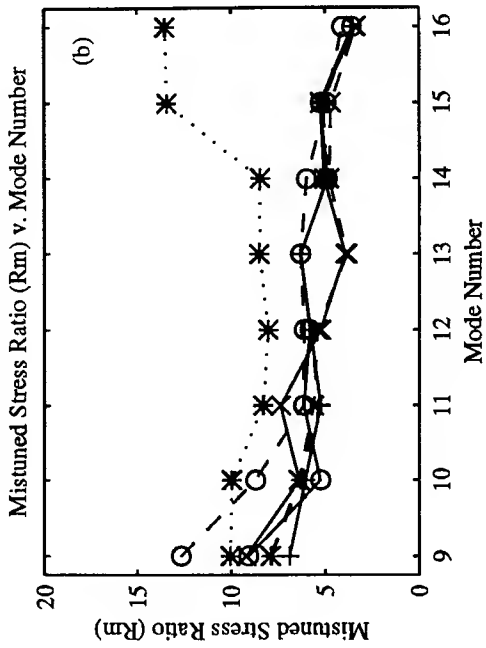
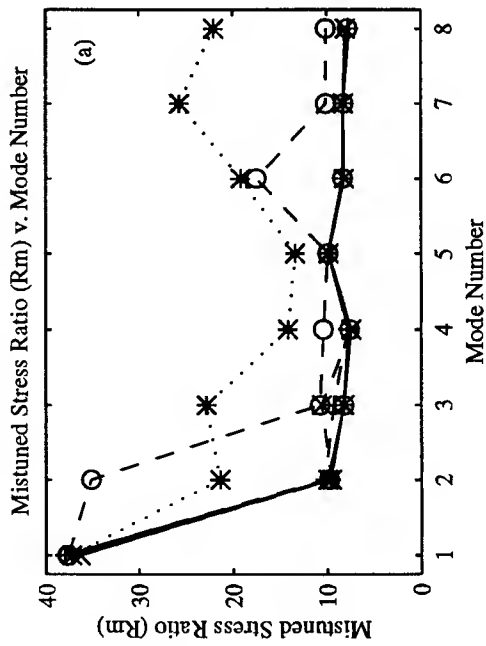


**Figure 11: Model 1 (baseline), eighth mode, 1% blade mass removed from blade 1.**



- = symmetric models
- = non-symmetric (mass added) models
- .. = non-symmetric (variable stiffness) models
- o = constant disk stiffness
- x = reduced interior stiffness
- + = reduced exterior stiffness
- \* = variable stiffness

Figure 12: (a) First modal group, (b) Second modal group, (c) Third modal group.



- = symmetric models
- = non-symmetric (mass added) models
- .. = non-symmetric (variable stiffness) models
- o = constant disk stiffness
- x = reduced interior stiffness
- + = reduced exterior stiffness
- \* = variable stiffness

Figure 13: Mistuning Case 1 (1% Mass added to blade 1): (a) First modal group, (b) Second modal group, (c) Third modal group.



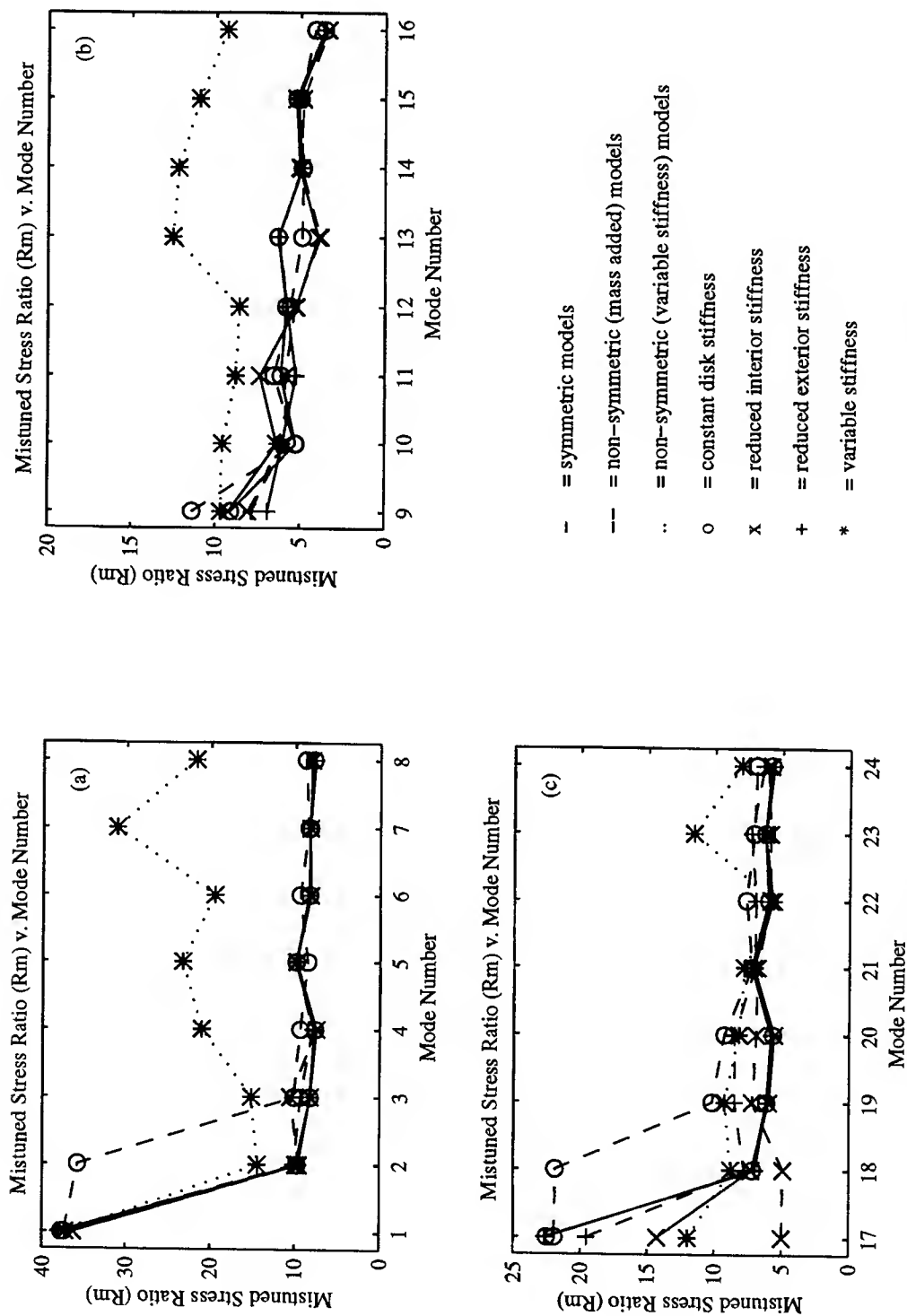
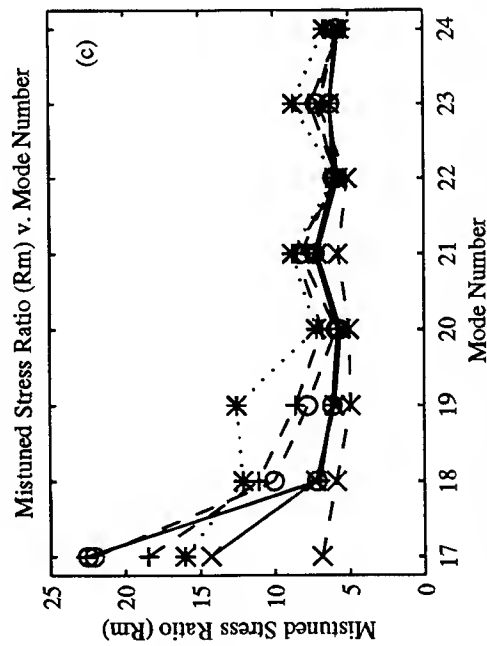
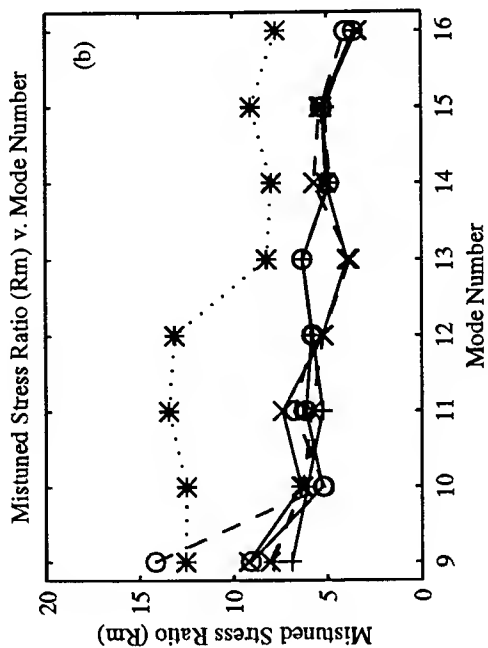
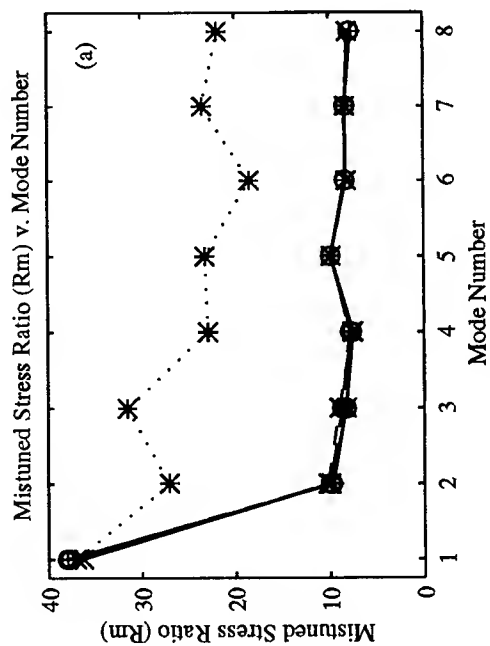


Figure 14: Mistuning Case 2 (1 % Mass added to blade 2): (a) First modal group, (b) Second modal group, (c) Third modal group.



- = symmetric models
- = non-symmetric (mass added) models
- .. = non-symmetric (variable stiffness) models
- o = constant disk stiffness
- x = reduced interior stiffness
- + = reduced exterior stiffness
- \* = variable stiffness

Figure 15: Mistuning Case 3 (1% Mass added to blade 3): (a) First modal group, (b) Second modal group, (c) Third modal group.

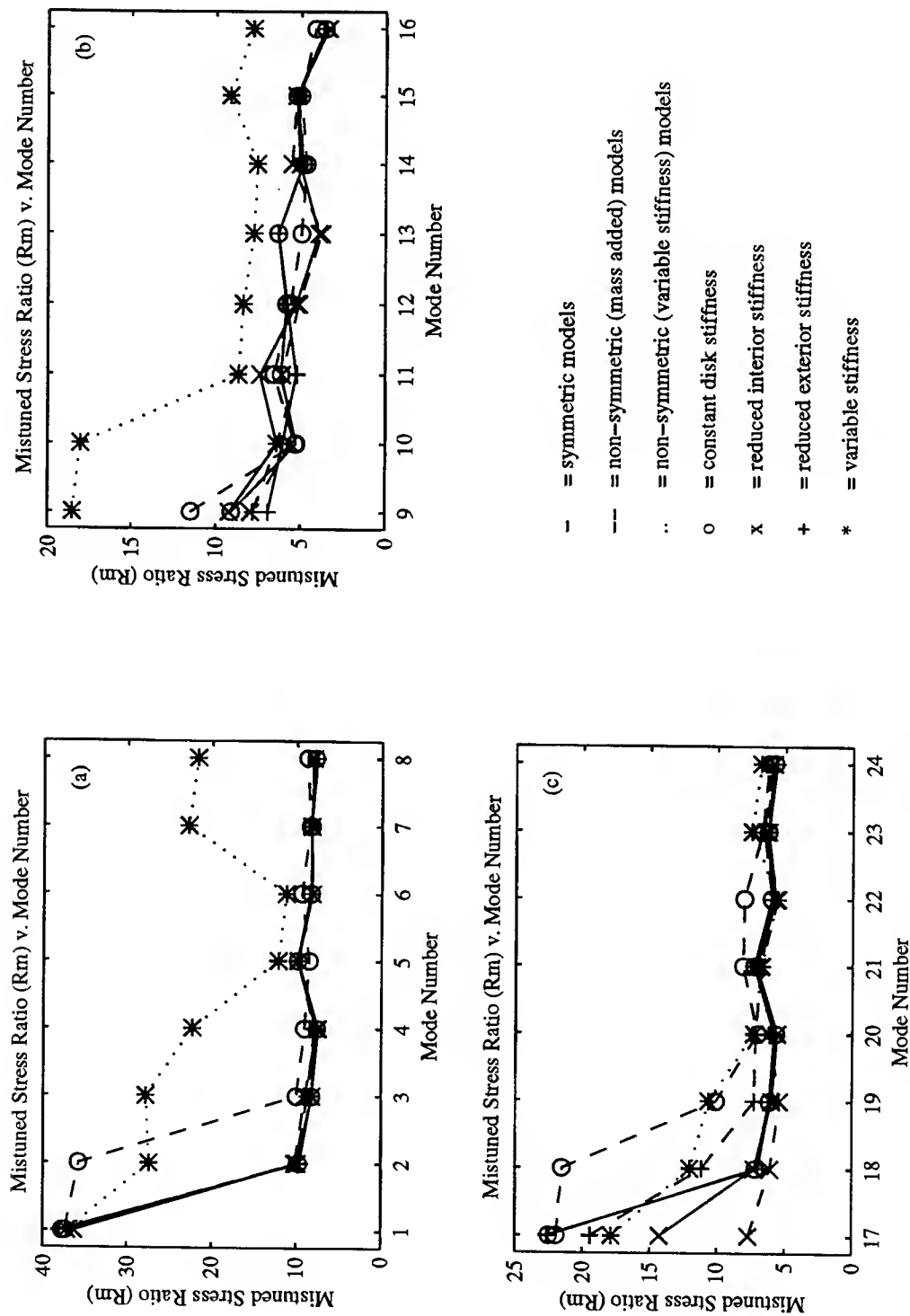


Figure 16: Mistuning Case 4 (1% Mass added to blade 4): (a) First modal group, (b) Second modal group, (c) Third modal group.

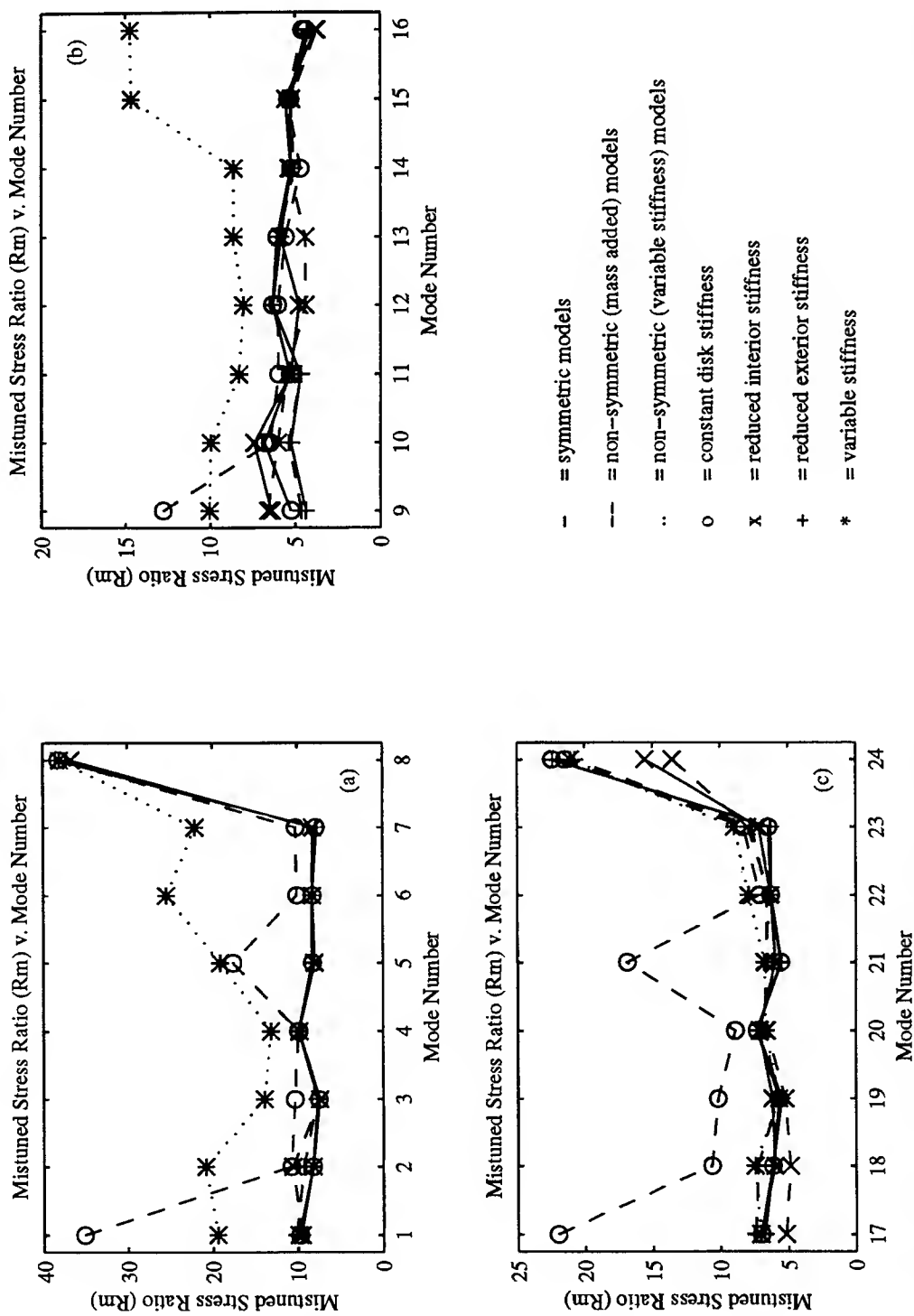


Figure 17: Mistuning Case 5 (1% Mass removed from blade 1): (a) First modal group, (b) Second modal group, (c) Third modal group.

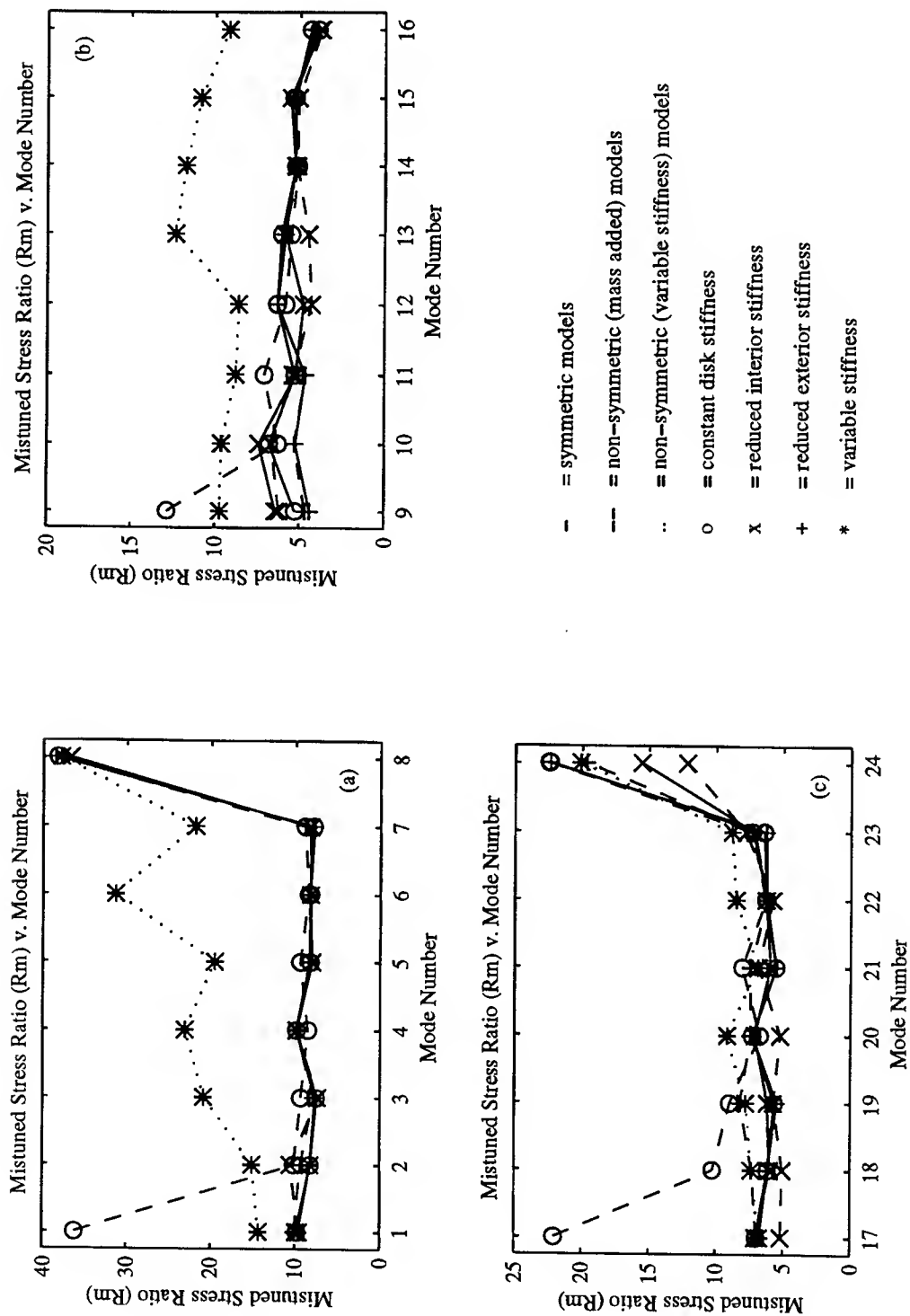
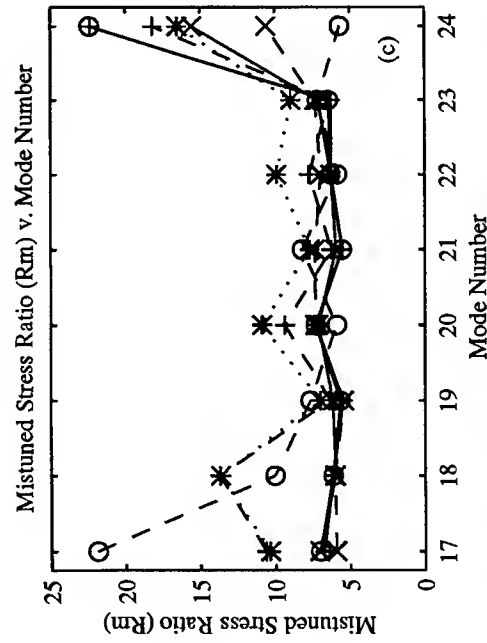
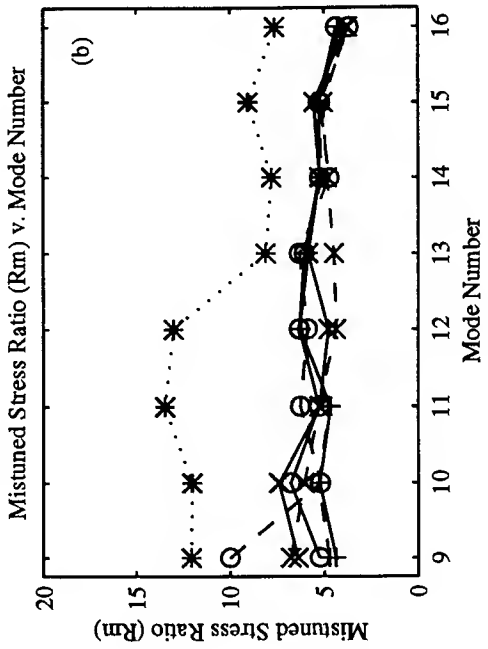
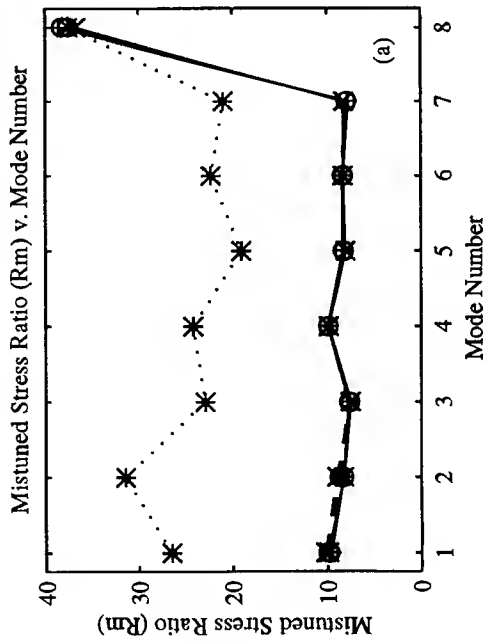


Figure 18: Mistuning Case 6 (1% Mass removed from blade 2): (a) First modal group, (b) Second modal group, (c) Third modal group.



- = symmetric models
- = non-symmetric (mass added) models
- .. = non-symmetric (variable stiffness) models
- o = constant disk stiffness
- x = reduced interior stiffness
- + = reduced exterior stiffness
- \* = variable stiffness

Figure 19: Mistuning Case 7 (1% Mass removed from blade 3): (a) First modal group, (b) Second modal group, (c) Third modal group.

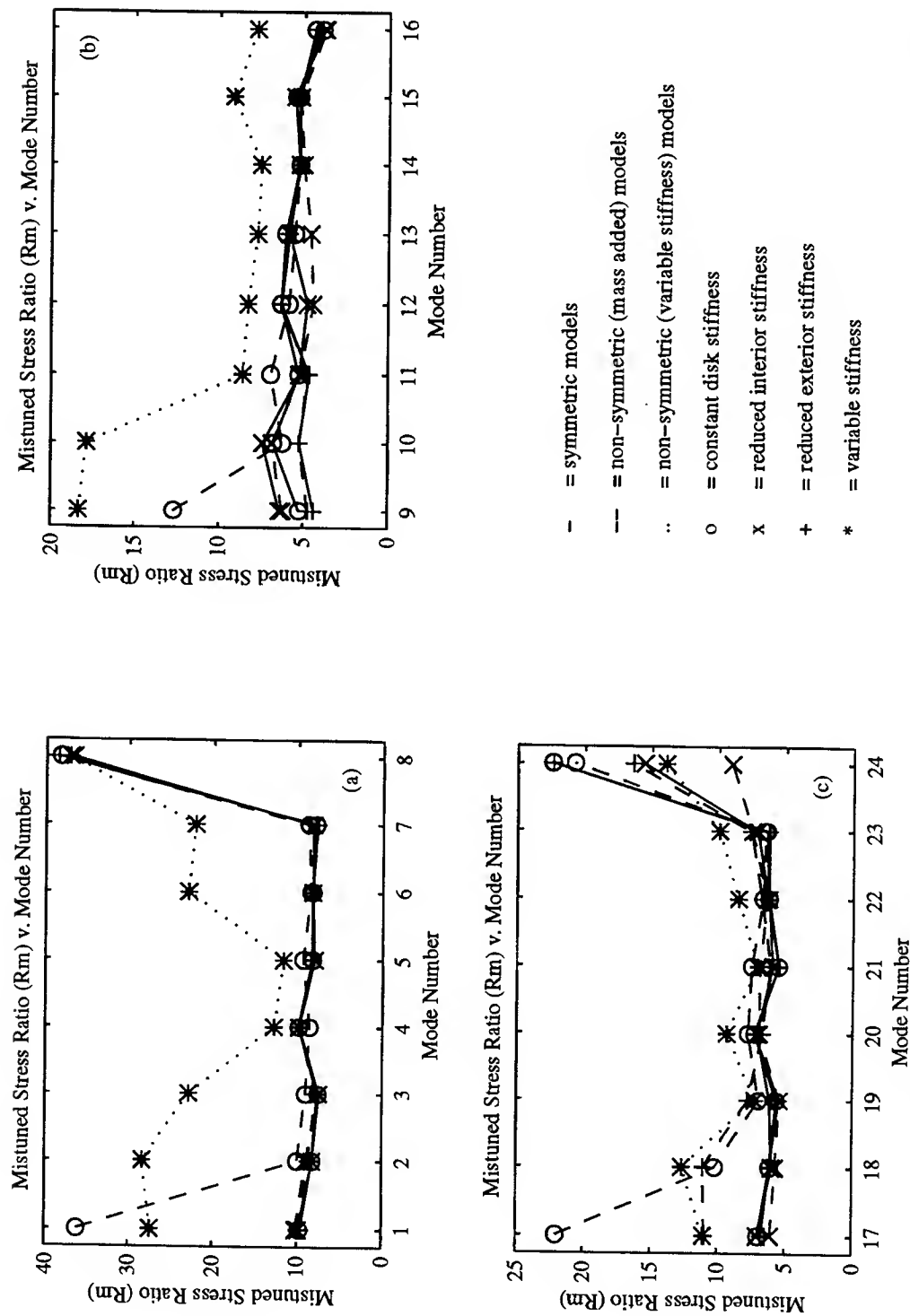
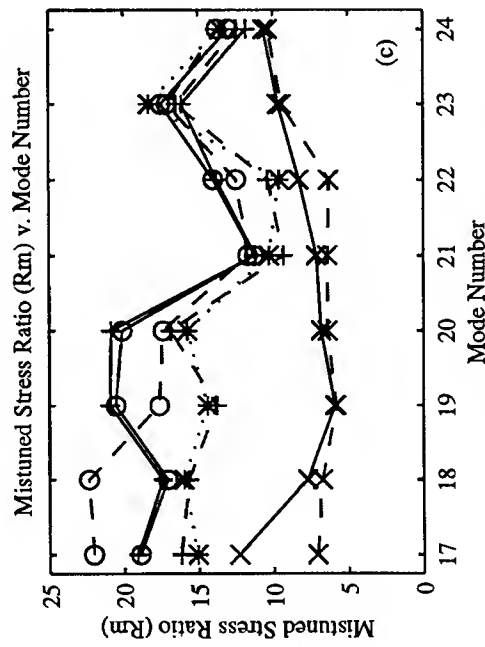
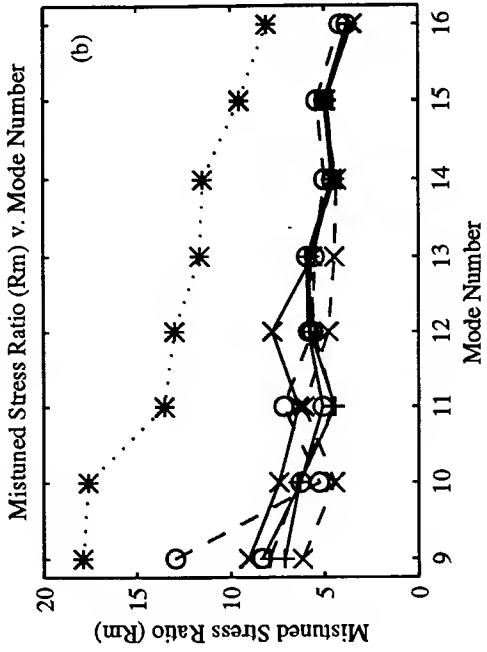
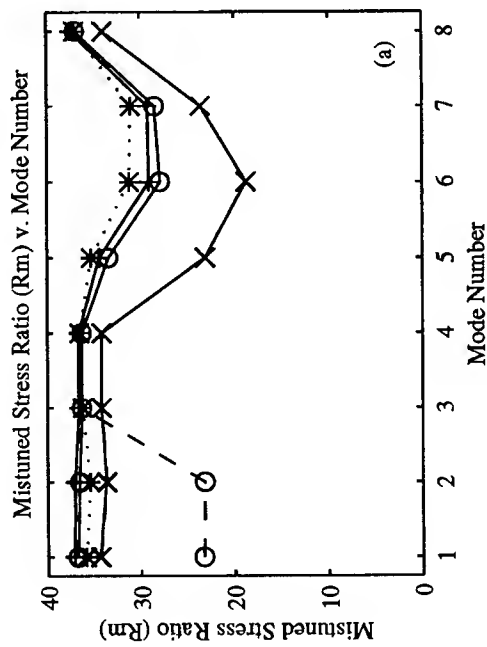


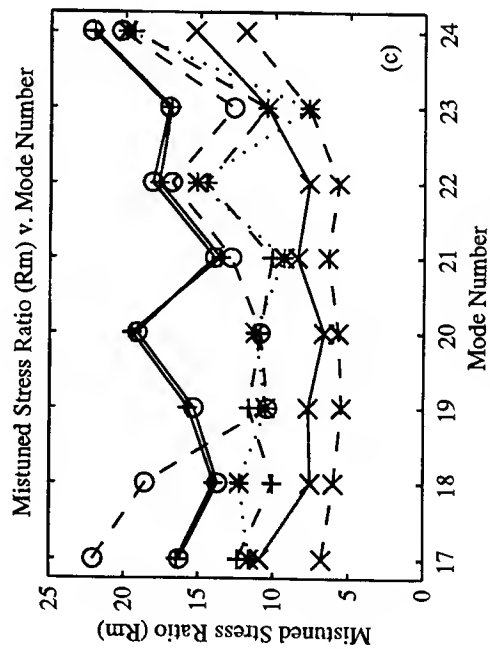
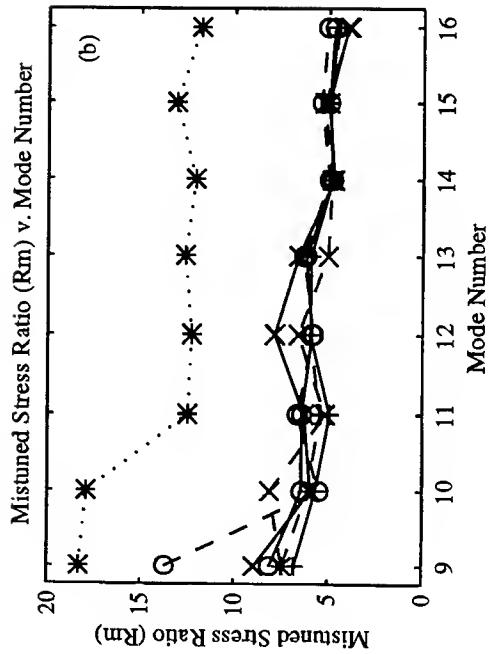
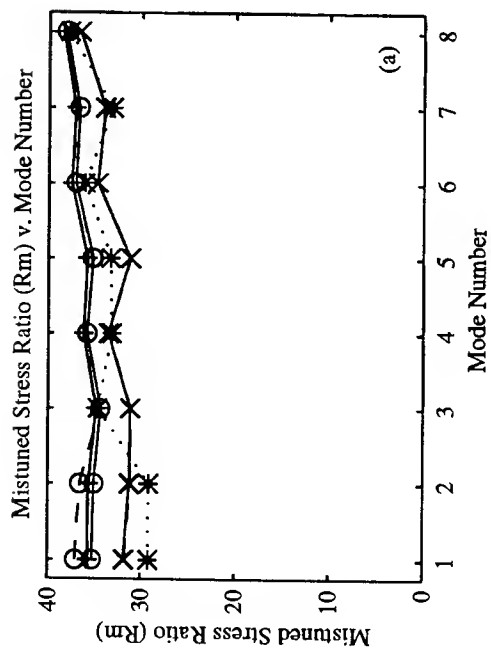
Figure 20: Mistuning Case 8 (1% Mass removed from blade 4): (a) First modal group, (b) Second modal group, (c) Third modal group.



- = symmetric models
- = non-symmetric (mass added) models
- .. = non-symmetric (variable stiffness) models
- o = constant disk stiffness
- x = reduced interior stiffness
- + = reduced exterior stiffness
- \* = variable stiffness

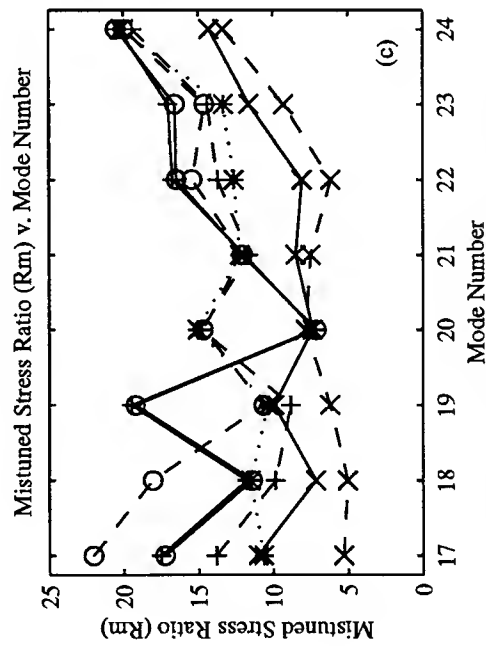
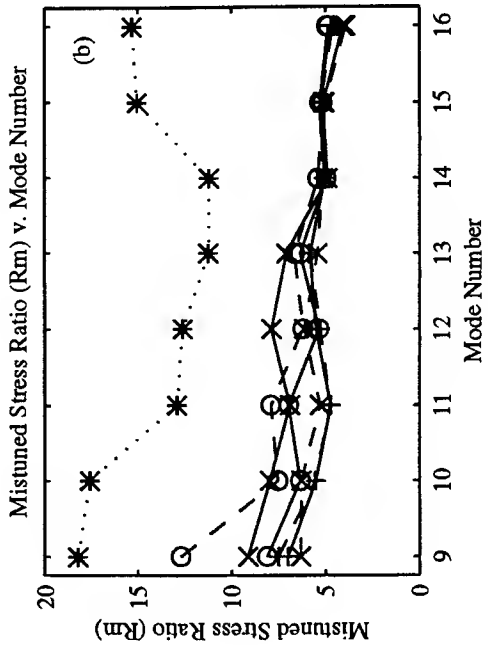
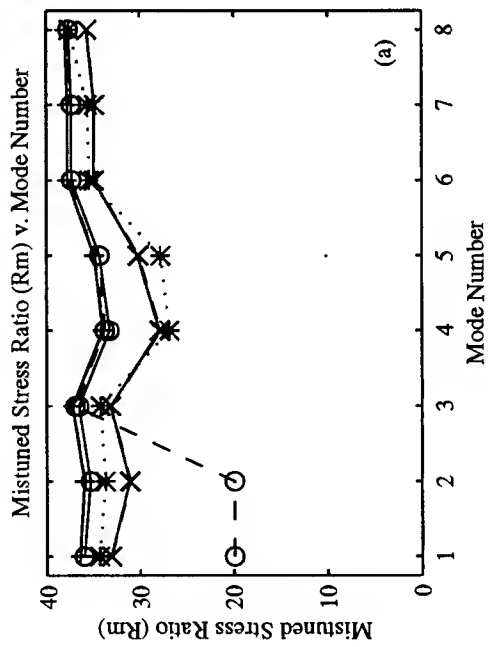
Figure 21: Mistuning Case 9 (Random pattern 1): (a) First modal group, (b) Second modal group, (c) Third modal group.





- = symmetric models
- = non-symmetric (mass added) models
- ... = non-symmetric (variable stiffness) models
- o = constant disk stiffness
- x = reduced interior stiffness
- +
- = reduced exterior stiffness
- \*
- = variable stiffness

Figure 22: Mistuning Case 10 (Random pattern 2): (a) First modal group, (b) Second modal group, (c) Third modal group.



- = symmetric models
- = non-symmetric (mass added) models
- .. = non-symmetric (variable stiffness) models
- o = constant disk stiffness
- x = reduced interior stiffness
- + = reduced exterior stiffness
- \* = variable stiffness

Figure 23: Mistuning Case 11 (Random pattern 3): (a) First modal group, (b) Second modal group, (c) Third modal group.

Mode	Tuned	Cases 1-4	Cases 5-8	Case 9	Case 10	Case 11
#	Hz	Hz	Hz	Hz	Hz	Hz
1	336.8	330.64	336.8	334.74	331.49	331.74
2	336.8	336.8	336.8	336.56	332.55	333.6
3	336.8	336.8	336.85	338.01	334.54	333.92
4	336.9	336.85	336.9	338.45	334.96	334.36
5	336.9	336.9	337	342.68	335.8	335.24
6	337.07	337	337.07	342.83	336.59	339.18
7	337.07	337.07	337.13	343.15	338.03	340.05
8	337.15	337.13	343.59	343.49	342.34	342.94
9	1411.4	1411.3	1411.4	1411.4	1411.3	1411.3
10	1411.4	1411.4	1411.4	1411.4	1411.3	1411.4
11	1411.4	1411.4	1411.4	1411.4	1411.4	1411.4
12	1411.5	1411.5	1411.5	1411.5	1411.5	1411.5
13	1411.5	1411.5	1411.5	1411.6	1411.5	1411.5
14	1411.8	1411.8	1411.8	1411.9	1411.8	1411.8
15	1411.8	1411.8	1411.9	1411.9	1411.8	1411.9
16	1412.0	1412	1412	1412	1412	1412
17	2066.1	2038.4	2066.5	2059.6	2040.9	2042.6
18	2066.7	2066.5	2066.8	2073.1	2051	2055.1
19	2066.7	2066.8	2069.5	2078.6	2057.8	2056.1
20	2072.8	2070.3	2072.8	2081	2061.3	2059
21	2072.8	2072.8	2076.9	2104.3	2070.6	2067.2
22	2081.4	2078.5	2081.6	2108.9	2072.5	2088.6
23	2081.4	2081.6	2083.7	2113.9	2082.8	2091.3
24	2084.6	2083.9	2114.1	2115.4	2106.7	2110.9

**Table 4: Frequency table for Model 1 (symmetric mass distribution, axisymmetric stiffness).**

Mode #	Tuned Hz	Case 1 Hz	Case 2 Hz	Case 3 Hz	Case 4 Hz	Case 5 Hz	Case 6 Hz	Case 7 Hz	Case 8 Hz	Case 9 Hz	Case 10 Hz	Case 11 Hz
1	335.03	330.64	330.64	328.86	330.64	335.03	335.03	336.8	335.03	334.59	329.7	331.65
2	336.8	335.03	335.03	336.8	335.03	336.8	336.8	336.8	336.8	334.81	332.54	331.85
3	336.8	336.8	336.8	336.8	336.8	336.83	336.81	336.85	336.82	338	334.54	333.92
4	336.85	336.83	336.81	336.85	336.82	336.89	336.87	336.89	336.87	338.45	334.96	334.36
5	336.89	336.89	336.87	336.89	336.87	336.96	336.95	337	336.95	342.68	335.8	335.24
6	337.01	336.96	336.96	337	336.96	337.02	337.06	337.07	337.06	342.83	336.59	339.18
7	337.07	337.03	337.06	337.07	337.06	337.11	337.12	337.13	337.12	343.15	338.03	340.04
8	337.13	337.11	337.12	337.13	337.12	343.59	343.59	341.56	343.58	343.48	342.33	342.94
9	1411.2	1411.2	1411.2	1411.1	1411.2	1411.2	1411.2	1411.3	1411.2	1411.2	1411.1	1411.2
10	1411.3	1411.3	1411.3	1411.3	1411.3	1411.4	1411.4	1411.3	1411.4	1411.4	1411.3	1411.3
11	1411.4	1411.4	1411.4	1411.4	1411.4	1411.4	1411.4	1411.4	1411.4	1411.4	1411.4	1411.4
12	1411.4	1411.4	1411.4	1411.4	1411.4	1411.5	1411.4	1411.4	1411.4	1411.5	1411.4	1411.4
13	1411.5	1411.5	1411.4	1411.5	1411.4	1411.5	1411.5	1411.5	1411.5	1411.5	1411.5	1411.5
14	1411.8	1411.8	1411.8	1411.8	1411.8	1411.8	1411.8	1411.8	1411.8	1411.8	1411.7	1411.7
15	1411.8	1411.8	1411.8	1411.8	1411.8	1411.8	1411.8	1411.8	1411.8	1411.9	1411.8	1411.8
16	1412	1412	1412	1412	1412	1412	1412	1412	1412	1412	1412	1412
17	1711.9	1711.9	1711.9	1695.7	1711.9	1711.9	1711.9	1729.3	1711.9	1710.9	1697.9	1703.3
18	2055.7	2035.3	2036	2055.7	2031.9	2056.5	2055.7	2055.7	2055.7	2053.9	2041.2	2040.7
19	2060.1	2057.2	2055.8	2060.1	2055.7	2060.6	2060.5	2060.1	2062.2	2068	2051.4	2049.7
20	2064.4	2060.9	2060.8	2064.4	2062.6	2066.5	2064.9	2064.5	2067	2076.1	2054.8	2051.9
21	2067.5	2066.8	2065	2067.5	2067.1	2070.4	2072	2067.5	2071.5	2099.7	2067.2	2064.1
22	2075.6	2073	2073.2	2075.5	2073.3	2076.4	2078.8	2075.7	2079.2	2104.7	2070.7	2081.6
23	2080.6	2077	2079.4	2080.6	2079.8	2082.3	2082.6	2080.6	2082.5	2112.5	2078.5	2085.2
24	2083	2082.4	2082.7	2083	2082.7	2112.7	2110.6	2083	2106.4	2114.2	2103.8	2109.7

Table 5: Frequency table for Model 2 (Non-symmetric mass distribution, axisymmetric stiffness).

Mode #	Tuned Hz	Cases 1-4 Hz	Cases 5-8 Hz	Case 9 Hz	Case 10 Hz	Case 11 Hz
1	336.39	330.39	336.39	334.47	331.22	331.47
2	336.39	336.39	336.39	336.32	332.32	333.36
3	336.39	336.39	336.51	337.74	334.26	333.67
4	336.63	336.52	336.63	338.19	334.7	334.09
5	336.63	336.63	336.82	342.42	335.58	335.02
6	336.98	336.84	336.98	342.51	336.34	338.94
7	336.98	336.98	337.07	342.97	337.8	339.79
8	337.1	337.07	343.34	343.24	342.09	342.7
9	1411.1	1411.1	1411.1	1411.1	1411.1	1411.1
10	1411.1	1411.1	1411.1	1411.2	1411.1	1411.1
11	1411.1	1411.1	1411.1	1411.2	1411.1	1411.1
12	1411.1	1411.1	1411.2	1411.2	1411.1	1411.2
13	1411.1	1411.2	1411.2	1411.3	1411.2	1411.2
14	1411.7	1411.6	1411.7	1411.7	1411.6	1411.6
15	1411.7	1411.7	1411.7	1411.7	1411.7	1411.7
16	1412	1411.9	1412	1412	1411.9	1412
17	2031.4	2012.7	2031.7	2030	2010.9	2012.4
18	2031.9	2031.7	2032.1	2047.2	2024.2	2025.9
19	2031.9	2032.1	2038.8	2050.2	2031.2	2030.8
20	2055.8	2049.2	2055.9	2061.1	2040.9	2044
21	2055.8	2055.9	2062.4	2081.2	2055.4	2058.1
22	2078.4	2072.3	2078.5	2097.2	2064.3	2070.8
23	2078.4	2078.5	2081.7	2105.9	2076.6	2081.6
24	2083.6	2082.4	2103.8	2110.5	2096.7	2102.5

**Table 6: Frequency table for Model 3 (Symmetric mass distribution, reduced interior stiffness).**

Mode #	Tuned Hz	Cases 1-4 Hz	Cases 5-8 Hz	Case 9 Hz	Case 10 Hz	Case 11 Hz
1	336.06	329.91	336.06	333.99	330.76	331.01
2	336.06	336.06	336.06	335.8	331.81	332.85
3	336.06	336.06	336.1	337.25	333.79	333.18
4	336.14	336.11	336.14	337.69	334.21	333.61
5	336.14	336.14	336.23	341.9	335.05	334.49
6	336.3	336.24	336.3	342.06	335.83	338.42
7	336.3	336.3	336.36	342.36	337.27	339.28
8	336.38	336.36	342.81	342.71	341.56	342.16
9	1409.5	1409.5	1409.5	1409.5	1409.4	1409.4
10	1409.5	1409.5	1409.5	1409.6	1409.5	1409.5
11	1409.5	1409.5	1409.6	1409.6	1409.5	1409.5
12	1409.8	1409.7	1409.8	1409.8	1409.7	1409.7
13	1409.8	1409.8	1409.8	1409.8	1409.8	1409.8
14	1410.1	1410.1	1410.1	1410.2	1410.1	1410.1
15	1410.1	1410.1	1410.2	1410.2	1410.1	1410.2
16	1410.3	1410.3	1410.3	1410.4	1410.3	1410.3
17	2058.2	2030.5	2058.4	2051.5	2033.1	2034.8
18	2058.8	2058.4	2058.8	2064.9	2043	2047
19	2058.8	2058.8	2061.3	2070.3	2049.8	2048.1
20	2064.3	2062	2064.3	2072.7	2053.3	2050.8
21	2064.3	2064.3	2068.6	2095.7	2062.3	2059.1
22	2073.2	2070.2	2073.2	2100.2	2064.4	2080.3
23	2073.2	2073.2	2075.7	2105.4	2074.5	2082.9
24	2076.7	2075.9	2105.5	2106.9	2098.2	2102.3

**Table 7: Frequency table for Model 4 (Symmetric mass distribution, reduced exterior stiffness).**

Mode #	Tuned Hz	Case 1 Hz	Case 2 Hz	Case 3 Hz	Case 4 Hz	Case 5 Hz	Case 6 Hz	Case 7 Hz	Case 8 Hz	Case 9 Hz	Case 10 Hz	Case 11 Hz
1	336.38	330.39	330.39	330.39	330.38	336.38	336.38	336.38	336.38	334.46	331.21	331.47
2	336.38	336.38	336.38	336.38	336.38	336.38	336.38	336.38	336.38	336.31	332.32	333.35
3	336.38	336.38	336.38	336.38	336.38	336.5	336.5	336.51	336.5	337.73	334.26	333.66
4	336.62	336.51	336.51	336.51	336.51	336.62	336.62	336.62	336.62	338.18	334.69	334.09
5	336.62	336.62	336.62	336.62	336.62	336.82	336.82	336.82	336.82	342.41	335.58	335.01
6	336.98	336.83	336.83	336.83	336.83	336.98	336.98	336.98	336.98	342.5	336.34	338.93
7	336.98	336.98	336.98	336.98	336.98	337.07	337.07	337.07	337.07	342.97	337.79	339.78
8	337.1	337.07	337.07	337.07	337.07	343.33	343.33	343.33	343.33	343.23	342.08	342.69
9	1411	1411	1411	1411	1411	1411	1411	1411	1411	1411	1411	1411
10	1411	1411	1411	1411	1411	1411	1411.1	1411	1411.1	1411.1	1411	1411
11	1411.1	1411.1	1411.1	1411.1	1411.1	1411.1	1411.1	1411.1	1411.1	1411.1	1411.1	1411.1
12	1411.1	1411.1	1411.1	1411.1	1411.1	1411.1	1411.1	1411.1	1411.1	1411.2	1411.1	1411.1
13	1411.2	1411.2	1411.2	1411.2	1411.2	1411.2	1411.2	1411.2	1411.2	1411.3	1411.2	1411.2
14	1411.6	1411.6	1411.6	1411.6	1411.6	1411.6	1411.6	1411.6	1411.6	1411.6	1411.6	1411.6
15	1411.6	1411.6	1411.6	1411.6	1411.6	1411.6	1411.6	1411.6	1411.6	1411.7	1411.6	1411.6
16	1412	1411.9	1411.9	1411.9	1411.9	1412	1412	1412	1412	1412	1411.9	1412
17	1946.5	1945.5	1945.7	1945.7	1939.3	1947	1946.9	1949.4	1949.8	1948.5	1936.8	1945.1
18	1964.7	1962.6	1962.8	1962.8	1960.5	1965.7	1965.5	1968.4	1969.6	1971.4	1959.3	1961.2
19	1997.1	1990.2	1988.4	1988.4	1995.9	2000.9	2001.5	1998.4	1998.2	2013	1993.2	1991.3
20	2032.6	2031.5	2023.3	2023.3	2027	2032.6	2039.2	2033.3	2037.4	2040.4	2026.2	2019.9
21	2042.8	2032.6	2042.3	2042.8	2042.8	2050	2044.9	2047.5	2042.8	2067.2	2038.5	2042.7
22	2075.3	2067.7	2068.2	2067.2	2067.8	2075.5	2075.5	2076.6	2076.5	2090.2	2060.8	2061.3
23	2076.8	2075.5	2075.6	2076.6	2076.6	2080.3	2080.1	2079.7	2079.4	2102.9	2071.9	2075.9
24	2082.6	2081.2	2081.2	2081.1	2081.1	2100.9	2098.5	2095.8	2093.3	2108.3	2091.4	2099.6

Table 8: Frequency table for Model 5 (Non-symmetric mass distribution, reduced interior stiffness).

Mode #	Tuned Hz	Case 1 Hz	Case 2 Hz	Case 3 Hz	Case 4 Hz	Case 5 Hz	Case 6 Hz	Case 7 Hz	Case 8 Hz	Case 9 Hz	Case 10 Hz	Case 11 Hz
1	336.05	329.91	329.91	329.9	329.9	336.05	336.05	336.05	336.05	333.99	330.75	331
2	336.06	336.05	336.05	336.05	336.05	336.06	336.06	336.06	336.06	335.79	331.8	332.85
3	336.06	336.06	336.06	336.06	336.06	336.1	336.1	336.1	336.1	337.24	333.79	333.17
4	336.14	336.1	336.1	336.1	336.1	336.14	336.14	336.14	336.14	337.69	334.21	333.61
5	336.14	336.14	336.14	336.14	336.14	336.23	336.23	336.23	336.23	341.9	335.04	334.48
6	336.3	336.23	336.23	336.23	336.23	336.3	336.3	336.3	336.3	342.05	335.83	338.41
7	336.3	336.3	336.3	336.3	336.3	336.35	336.35	336.35	336.35	342.36	337.27	339.27
8	336.37	336.36	336.36	336.36	336.36	342.81	342.8	342.8	342.8	342.7	341.56	342.16
9	1409.5	1409.4	1409.4	1409.4	1409.4	1409.5	1409.5	1409.5	1409.5	1409.5	1409.4	1409.4
10	1409.5	1409.5	1409.5	1409.5	1409.5	1409.5	1409.5	1409.5	1409.5	1409.6	1409.5	1409.5
11	1409.5	1409.5	1409.5	1409.5	1409.5	1409.5	1409.5	1409.5	1409.5	1409.6	1409.5	1409.5
12	1409.7	1409.7	1409.7	1409.7	1409.7	1409.7	1409.7	1409.7	1409.7	1409.8	1409.7	1409.7
13	1409.7	1409.7	1409.7	1409.7	1409.7	1409.8	1409.8	1409.8	1409.8	1409.8	1409.7	1409.7
14	1410.1	1410.1	1410.1	1410.1	1410.1	1410.1	1410.1	1410.1	1410.1	1410.2	1410.1	1410.1
15	1410.1	1410.1	1410.1	1410.1	1410.1	1410.2	1410.2	1410.2	1410.2	1410.2	1410.1	1410.2
16	1410.3	1410.3	1410.3	1410.3	1410.3	1410.3	1410.3	1410.3	1410.3	1410.4	1410.3	1410.3
17	2041.3	2025.8	2025.8	2019.3	2016.3	2041.5	2041.5	2042	2042	2035	2018.3	2026.3
18	2042.7	2041.7	2041.6	2042	2042	2043.3	2043.1	2046.4	2048.4	2053.5	2034.4	2035.4
19	2050.6	2044.4	2043.5	2048.8	2049.8	2052.3	2052.2	2051.9	2051.9	2057.3	2038.8	2040.3
20	2055.1	2053.8	2052.7	2053.2	2053.3	2055.1	2058.2	2056.2	2059	2065.2	2045.2	2041.4
21	2059	2055.2	2058.5	2057	2059	2063.3	2062.8	2062.6	2060.9	2089.4	2057.5	2054.6
22	2069.2	2065.9	2066.2	2065.5	2066	2069.4	2069.7	2071	2070.8	2093.6	2061.6	2069.6
23	2071.2	2069.5	2069.8	2071	2070.9	2073.6	2073.5	2073.3	2073.2	2103.1	2067.4	2074.2
24	2075	2074	2074	2074	2074	2103.3	2100.8	2097.7	2094.3	2104.9	2093.6	2100.3

Table 9: Frequency table for Model 6 (Non-symmetric mass distribution, reduced exterior stiffness).

Mode #	Tuned Hz	Case 1 Hz	Case 2 Hz	Case 3 Hz	Case 4 Hz	Case 5 Hz	Case 6 Hz	Case 7 Hz	Case 8 Hz	Case 9 Hz	Case 10 Hz	Case 11 Hz
1	335.57	330.55	330.36	329.93	329.38	335.58	335.57	335.58	335.58	333.45	330.73	331.45
2	335.58	335.58	335.57	335.58	335.58	335.59	335.58	335.62	336.16	335.83	331.32	332.88
3	336.2	335.59	335.58	335.63	336.17	336.2	336.2	336.2	336.2	336.7	334.22	333.63
4	336.21	336.2	336.2	336.2	336.2	336.21	336.24	336.6	336.62	338.15	334.44	333.9
5	336.62	336.21	336.25	336.6	336.62	336.63	336.63	336.63	336.63	341.93	335.51	334.33
6	336.64	336.63	336.63	336.63	336.63	336.69	336.86	336.91	336.9	342.53	336.5	338.45
7	336.92	336.69	336.87	336.91	336.91	336.92	336.92	336.92	336.92	343.05	336.74	338.73
8	336.92	336.92	336.92	336.92	336.92	343.5	343.29	342.83	342.25	343.39	342.04	342.85
9	1408.8	1408.8	1408.8	1408.8	1408.7	1408.8	1408.8	1408.8	1408.8	1408.7	1408.7	1408.7
10	1408.8	1408.8	1408.8	1408.8	1408.8	1408.8	1408.8	1408.8	1408.9	1408.8	1408.8	1408.8
11	1410.2	1410.2	1410.2	1410.1	1410.2	1410.2	1410.2	1410.2	1410.2	1410.2	1410.1	1410.1
12	1410.2	1410.2	1410.2	1410.2	1410.2	1410.2	1410.2	1410.3	1410.2	1410.3	1410.2	1410.2
13	1411	1411	1410.9	1411	1411	1411	1411	1411	1411	1411	1411	1410.9
14	1411	1411	1411	1411	1411	1411	1411.1	1411	1411	1411.1	1411.1	1411
15	1411.6	1411.5	1411.6	1411.6	1411.6	1411.6	1411.6	1411.6	1411.6	1411.7	1411.5	1411.5
16	1411.6	1411.6	1411.6	1411.6	1411.6	1411.7	1411.6	1411.6	1411.6	1411.7	1411.6	1411.7
17	2032.4	2027.5	2027.4	2014.1	2009.3	2033	2032.9	2034.2	2034.3	2027.2	2010.6	2025.2
18	2036.2	2034.6	2034.4	2034.7	2034.6	2036.8	2036.6	2040.9	2045.4	2046.5	2032.3	2030.1
19	2054.6	2040.6	2038	2048.5	2052	2056.1	2055.8	2055.5	2055.5	2057.8	2034.9	2040.9
20	2057.1	2056.9	2056.1	2055.8	2055.9	2057.5	2060.1	2061.4	2063.7	2070.5	2049.8	2046.6
21	2064	2058.3	2061.6	2063	2064.1	2067.6	2067.4	2066.5	2064.2	2093.5	2061.5	2059.7
22	2072	2069.6	2069.9	2069.2	2069.6	2073.4	2073.4	2076.5	2076.5	2097.8	2067.5	2068.3
23	2077.7	2074.3	2074.5	2077.3	2077.5	2079.1	2079	2078.6	2078.3	2109.5	2070.6	2075.4
24	2080	2079.2	2079.2	2079.2	2079.2	2109.7	2108	2098.7	2094.1	2111.5	2100.6	2106.5

Table 10: Frequency table for Model 7 (Symmetric mass distribution, non-symmetric stiffness distribution).



## References

- [1] Avitabile, P., O'Callahan, J.C., and Milani, J., "Comparison of System Characteristics Using Various Model Reduction Techniques," *Seventh International Modal Analysis Conference*, Las Vegas, Nevada, Feb. 1989.
- [2] Basu, P., and Griffin, J.H., "The Effect of Limiting Aerodynamic and Structural Coupling in Models of Mistuned Bladed Disk Vibration," *Journal of Vibration, Acoustics, Stress, and Reliability in Design*, Vol. 108, 1986, p. 132-139.
- [3] Bendiksen, O.O., "Flutter of Mistuned Turbomachinery Rotors," *Journal of Engineering for Gas Turbines and Power*, Vol. 106, pp. 25-33.
- [4] Blair, J.J., Personal Communication, July 1995.
- [5] Campbell, W., "The Protection of Steam-Turbine Disk Wheels From Axial Vibration", *ASME Transactions*, Vol. 46, No. 1920, 1924, pp. 31-160.
- [6] Craig, Jr., R.R., *Structural Dynamics*, Wiley, New York, 1981, pp. 467-494.
- [7] Craig, Jr., R.R., and Chung, Y., "Generalized Substructure Coupling Procedure for Damped Systems," *AIAA Journal*, Vol. 20, March 1982, pp. 442-444.
- [8] Craig, Jr., R.R., "A Review of Time-Domain and Frequency-Domain Component-Mode Synthesis Methods," *Journal of Modal Analysis*, April 1987, pp. 59-72.
- [9] Craig, Jr., R.R., and Hale, A.L., "Block-Krylov Component Synthesis Method for Structural Model Reduction," *Journal of Guidance Control and Dynamics*, Vol. 11, Nov.-Dec. 1988, pp. 562-570.
- [10] Crawley, E.F., and Mokadam, D.R., "Stagger Angle Dependence of Inertial and Elastic Coupling in Bladed Disks," *Journal of Vibration Acoustics, Stress, and Reliability in Design*, Vol. 106, 1984, pp. 181-189.
- [11] Dye, R.C.F., and Henry, T.A., "Vibration Amplitudes of Compressor Blades Resulting from Scatter in Natural Frequencies," *Journal of Engineering for Power*, Vol. 91, 1969, pp. 182-188.
- [12] Ewins, D.J., "The Effect of Detuning upon the Forced Vibrations of Bladed Disks," *Journal of Sound and Vibration*, Vol. 9, No. 1, 1969, pp. 65-79.
- [13] Ewins, D.J., "Vibration Characteristics of Bladed Disk Assemblies," *Journal of Mechanical Engineering Science*, Vol. 15, No. 3, 1973, pp. 165-186.
- [14] Ewins, D.J., and Han, Z.S., "Resonant Vibration Levels of a Mistuned Bladed Disk," *Journal of Vibration, Acoustics, Stress, and Reliability in Design*, Vol. 106, April 1984, pp. 211-217.
- [15] Gawronski, W., and Williams, T., "Model Reduction for Flexible Space Structures," *Journal of Guidance Control and Dynamics*, Vol. 14, Jan.-Feb. 1991, pp. 68-76.
- [16] Griffin, J.H., "On Predicting the Resonant Response of Bladed Disk Assemblies," *Journal of Engineering for Gas Turbines and Power*, Vol. 110, Jan. 1988, pp. 45-50.
- [17] Griffin, J.H., and Hoosac, T.M., "Model Development and Statistical Investigation of Turbine Blade Mistuning," *Journal of Vibration, Acoustics, Stress, and Reliability in Design*, Vol. 106, April 1984, pp. 204-210.
- [18] Hurty, W.C., "Dynamic Analysis of Structural Systems Using Component Modes," *AIAA Journal*, Vol. 3, April 1965, pp. 678-685.
- [19] Irretier, H., "Spectral Analysis of Mistuned Bladed Disk Assemblies by Component Mode Synthesis," *Proceedings of the Ninth Conference on Mechanical Vibration and Noise of the Design and Production Engineering Technical Conferences*, ASME, 1983, pp. 115-125.

- [20] Kaza, K.R.V., and Kielb, R.E., "Effects of Mistuning on Bending-Torsion Flutter and Response of a Mistuned Cascade in Incompressible Flow," *AIAA Journal*, Vol. 20, No. 8, 1982, pp. 1120-1127.
- [21] Kaza, K.R.V., and Kielb, R.E., "Flutter of Turbofan Rotors with Mistuned Blades," *AIAA Journal*, Vol. 22, No. 11, Nov. 1984, pp. 1618-1625.
- [22] Kielb, R.E., Personal Communications, August 1995.
- [23] Kielb, R.E., and Kaza, K.R.V., "Aeroelastic Characteristics of a Cascade of Mistuned Blades in Subsonic and Supersonic Flows," *Journal of Vibration, Acoustics, Stress, and Reliability in Design*, Vol. 105, Oct. 1983, pp. 425-433.
- [24] Kienholz, D.A., and Smith, K.E., "Admittance Modeling: Frequency Domain, Physical Coordinate Methods for Multi-Component Systems," CSA Engineering Report, Feb. 1994, pp. 608-614.
- [25] MacNeal, R.H., "A Hybrid Method of Component Mode Synthesis," *Computers and Structures*, Vol. 1, 1971, pp. 581-601.
- [26] Mead, D.J., "Wave Propagation and Natural Modes in Periodic Systems: Mono-Coupled Systems," *Journal of Sound and Vibration*, Vol. 40, No. 1, 1975, pp. 1-18.
- [27] Menq, C.-H., Griffin, J.H., and Bielak, J., "The Forced Response of Shrouded Fan Stages," *Journal of Vibration, Acoustics, Stress, and Reliability in Design*, Vol. 108, Jan. 1986, pp. 50-55.
- [28] Minas, C., and Kodiyalam, S., "Vibration Analysis of Bladed Disc Assemblies," *International Modal Analysis Conference*, 1995.
- [29] Murthy, D.V., Pierre, C., and Ottarsson, G., "Efficient Design Constraint Accounting for Mistuning Effects in Engine Rotors," *AIAA Journal*, Vol. 33, No. 5, 1994, pp. 960-962.
- [30] Muszyska, A., and Jones, D.I.G., "A Parametric Study of Dynamic Response of a Discrete Model of Turbomachinery Bladed Disc," *Transactions of the ASME*, Vol. 105, 1983, pp. 434-443.
- [31] Natsiavas, S., "Mode Localization and Frequency Veering In A Non-conservative Mechanical System With Dissimilar Components," *Journal of Sound and Vibration*, Vol. 165, 1993, pp. 137-147.
- [32] O'Callahan, J.C., Avitabile, P.A., Riemer, R., "System Equivalent Reduction Expansion Process (SEREP)," *Seventh International Modal Analysis Conference*, Las Vegas, Nevada, Feb. 1989.
- [33] O'Callahan, J.C., "A Procedure for an Improved Reduced System (IRS) Model," *Seventh International Modal Analysis Conference*, Las Vegas, Nevada, Feb. 1989.
- [34] Ottarsson, G., and Pierre, C., "A Transfer Matrix Approach to Vibration Localization in Mistuned Blade Assemblies," *Proceedings of the International Gas Turbine and Aeroengine Congress and Exposition*, Cincinnati, Ohio, 1993, ASME 93-GT-115, pp. 1-20.
- [35] Petrov, E.P., "Analysis and Optimal Control of Stress Amplitudes Upon Forced Vibration of Turbomachine Impellers with Mistuning," *International Union of Theoretical and Applied Mechanics Symposium on The Active Control of Vibration*, Sept. 5, 1994, University of Bath, UK, pp. 189-196.
- [36] Petrov, E.P., "Large-Scale Finite Element Models of Blade-Shroud and Blade-Disk Joints and Condensation Technique for Vibration Analysis of Turbomachine Impellers," *Proceedings of the 7th World Congress on Finite Element Methods: "FEM: Today and the Future"*, Monte-Carlo, 1993, pp. 507-513.
- [37] Pierre, C., and Murthy, D.V., "Aeroelastic Modal Characteristics of Mistuned Blade Assemblies: Mode Localization and Loss of Eigenstructure," *AIAA Journal*, Vol. 30, No. 10, 1992, pp. 2483-2496.
- [38] Pierre, C., Smith, T.E., and Murthy, D.V., "Localization of Aeroelastic Modes in Mistuned High-Energy Turbines," *Journal of Propulsion and Power*, Vol. 10, 1994, pp. 318-328.
- [39] Rao, J. S., *Turbomachine Blade Vibration*, Wiley, New York, 1991.

- [40] Su, T., and Craig, R.R., "Krylov Model Reduction Algorithm for Undamped Structural Dynamics," *Journal of Guidance Control and Dynamics*, Vol. 14, Nov.-Dec. 1991, pp. 1311-1313.
- [41] Swaminadham, M., Soni, M.L., Stange, W.A., and Reed, J.D., "On Model Generation and Modal Analysis of Flexible Bladed-Disc Assemblies," *Bladed Disk Assemblies*, ASME Vibration Conference, Cambridge, MA, Sept. 27-30, 1987, pp. 49.
- [42] Vakakis, A.F., "Non-Similar Normal Oscillations in a Strongly Non-Linear Discrete System," *Journal of Sound and Vibration*, Vol. 158, 1992, pp. 341-361.
- [43] Valero, N.A., and Bendiksen, O.O., "Vibration Characteristics of Mistuned Shrouded Blade Assemblies," *Journal of Engineering for Gas Turbines and Power*, Vol. 108, 1986, pp. 293-299.
- [44] Wei, S.T., and Pierre, C., "Localization Phenomena in Mistuned Assemblies with Cyclic Symmetry Part 1: Free Vibrations," *Journal of Vibration, Acoustics, Stress, and Reliability in Design*, Vol. 110, 1988, pp. 429-438.
- [45] Whitehead, D.S., "Effects of Mistuning on the Forced Vibration of Blades with Mechanical Coupling," *Journal of Mechanical Engineering Science*, Vol. 18, No. 6, 1976.
- [46] Yang, M.-T., and Griffin, J.H., "A Reduced Order Approach for The Vibration of Mistuned Bladed Disk Assemblies," *International Gas Turbine and Aeroengine Congress & Exposition*, 1995, ASME Paper No. 95-GT-454.

**GROWTH OF SILICON CARBIDE THIN FILMS  
BY  
MOLECULAR BEAM EPITAXY**

*Dr. John Chen  
Mr. Ronald Birkhahn  
and  
Prof. Andrew J. Steckl, Principal Investigator*

Nanoelectronics Laboratory  
University of Cincinnati  
899 Rhodes Hall  
P. O. Box 210030  
Cincinnati, OH 45221-0030  
e-mail: a.steckl@uc.edu

Final Report for:  
1996 Summer Research Extension Program  
Wright Laboratory

Sponsored by:  
Air Force Office of Scientific Research  
Bolling Air Force Base, DC

and

Wright Laboratory  
Wright-Patterson Air Force Base, OH

February 1997

# **GROWTH OF SILICON CARBIDE THIN FILMS BY MOLECULAR BEAM EPITAXY**

*Dr. John Chen  
Mr. Ronald Birkhahn  
and  
Prof. Andrew J. Steckl, Principal Investigator*

## Abstract

A system for the molecular beam epitaxy (MBE) of SiC thin films was installed in the Nanoelectronics Laboratory at University of Cincinnati. The MBE system combines several gas and solid sources. Preliminary results of SiC heteroepitaxial growth by MBE were obtained. We have demonstrated the MBE growth of SiC on Si and semiconductor-on-insulator (SOI) substrates. SiC growth was obtained by either propane carbonization of the Si surface or by the pyrolysis of silacyclobutane (SCB), or the sequential use of carbonization and pyrolysis. At a growth temperature of 800 °C, initial experiments indicate a SiC growth rate  $\sim 0.1 \text{ \AA/s}$ . The crystallinity of the film surface was investigated using reflection high electron energy diffraction (RHEED). Films grown under certain conditions produce RHEED patterns indicating crystalline cubic (3C) SiC, while under other conditions a RHEED pattern is observed indicating a combination of crystalline and poly-crystalline SiC. The thickness and composition of the SiC films was analyzed by secondary ion mass spectrometry (SIMS). SIMS depth profiles indicate that no (or very low) contamination from N, O, and B was found in the films.

## 1. INTRODUCTION

Silicon carbide (SiC) is a wide band gap semiconductor with high thermal stability and conductivity, high breakdown voltage, etc. Among many SiC polytypes, 3C-SiC (or  $\beta$ -SiC, the only cubic structure) is the most promising candidate for higher power, higher temperature and higher frequency electronic devices because of its high electron mobility ( $1000 \text{ cm}^2/\text{V s}$ )[1], high saturated drift velocity (above  $10^7 \text{ cm/s}$ )[2]. Since bulk crystals of 3C-SiC are very expensive and their size is very small ( $\sim 3 - 5 \text{ mm}$ ), heteroepitaxial growth on Si substrate has become an alternative method for growing SiC. Due to the large lattice constant mismatch (20%) and the thermal expansion coefficient difference (8%) between 3C-SiC and Si, considerable effort was devoted to improving the film quality, leading to a two-step epitaxial growth. The first step is carbonization of Si substrate to relieve the strain between SiC and Si, but this SiC layer is usually too thin (a few hundred Å) to fabricate devices. The second step is essential growth of SiC on SiC by introducing both Si and C precursors. Currently, CVD is widely employed to grow epitaxial 3C-SiC on Si. Since the epitaxial growth temperature in CVD is generally higher than  $1200^\circ\text{C}$ , which can cause deterioration of the film quality and redistribution of the dopants, reduction of growth temperatures must be realized in order to fabricate the SiC device. Molecular Beam Epitaxy (MBE) is a promising methods to reduce 3C-SiC growth temperatures.

The work described in this report was partially sponsored by Research and Development Laboratories as a Supplemental Research Extension Program from AFOSR. The following sections of this report consist of: a description of the Riber 32 MBE system recently installed at the University of Cincinnati; a brief summary of previous work on growing SiC with MBE; preliminary results on MBE growth of 3C-SiC on Si at Cincinnati.

## 2. SYSTEM DESCRIPTION

The MBE system consists of these primary components: a control computer (1), control electronics and power supplies(2), gas cabinet (3), growth chamber (4), a load lock chamber (5), pumping systems. A schematic of the system is shown in Fig. 1.

## 2.1 LOAD LOCK

The load lock (5) serves as a buffer chamber between atmosphere and the growth

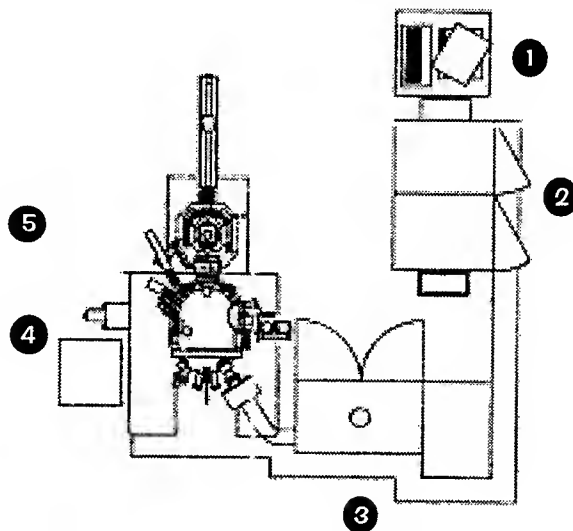


Fig. 1. Schematic of University of Cincinnati MBE system from Riber.

chamber which maintains UHV conditions. It contains a magnetically coupled sample transfer rod and a four-sample storage stage with one outgassing station that can reach temperature up to 800°C. The stage can move in XYZ directions and rotate 360 °C. The load lock is pumped by an ion pump (200 l/s) supplemented with a titanium sublimation pump (1000 l/s). The base pressure of the chamber can reach  $1 \times 10^{-10}$  Torr after a 48 hour bakeout at 200°C. One UHV port for venting the chamber with UHP nitrogen and for rough pumping is located just below the ion pressure gauge. There is a gate valve to isolate the load lock chamber from the ion pump during venting and roughing, and another gate valve to isolate the load lock from the growth chamber. On one side of the load lock chamber is a back door for loading and unloading samples. The load lock has additional ports for future expansion with a surface analysis chamber. A photo of the UC MBE-32 load lock is shown in Fig. 2.

## 2.2 ROUGH PUMPING

The rough pumping of the load lock and the growth chamber is done with a “rough

pump cart" which consists of two LN<sub>2</sub>-cooled sorption pumps in parallel with a small mechanical pump (Fig. 3). The sorption pumps are first cooled to 77K and the mechanical



Fig. 2. Photo of load lock with control electronics in background.

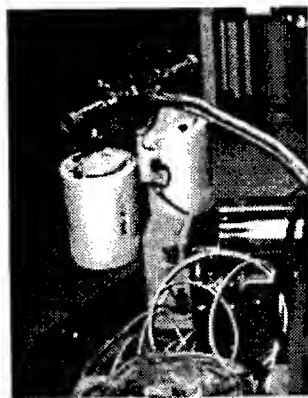


Fig. 3. Rough pump cart with mechanical and sorption pumps.

pump is allowed to rough out the load lock to 150 Torr. Next, the two sorption pumps



are opened and closed in series until the pressure reaches  $5 \times 10^{-2}$  Torr and  $1 \times 10^{-3}$  Torr, respectively. At that point, the chamber is isolated and the gate valve is opened to the main pumping system.

### 2.3 GROWTH CHAMBER

The growth chamber (4) (shown in Fig. 4) contains a sample manipulator, a cell panel where the molecular sources are located, analysis tools (RHEED, RGA),  $\text{LN}_2$  cryoshrouds, and UHV pumping. The transfer rod places the sample directly onto a xyz-controllable and rotatable manipulator (Fig. 5) that has heating capability to  $1200^\circ\text{C}$  via a specially designed heater from Karl Eberl (Stuttgart, Germany). UHV pumping consists of a CTI CT8 cryopump and a titanium sublimator. Pressures can go as low as  $5 \times 10^{-11}$  Torr after a 48 hour bakeout and  $1.2 \times 10^{-11}$  Torr with the cryopanel cooled. The cryopanel consists of three  $\text{LN}_2$ -filled shrouds for cooling the main chamber, the well surrounding the pumping systems, the cell panel to protect the sources from stray contaminants. Two ion gauges monitor the pressure in the system, one close to the pumping well and the second at the rear of the substrate holder to calibrate the fluxes from the cells incident on the sample.

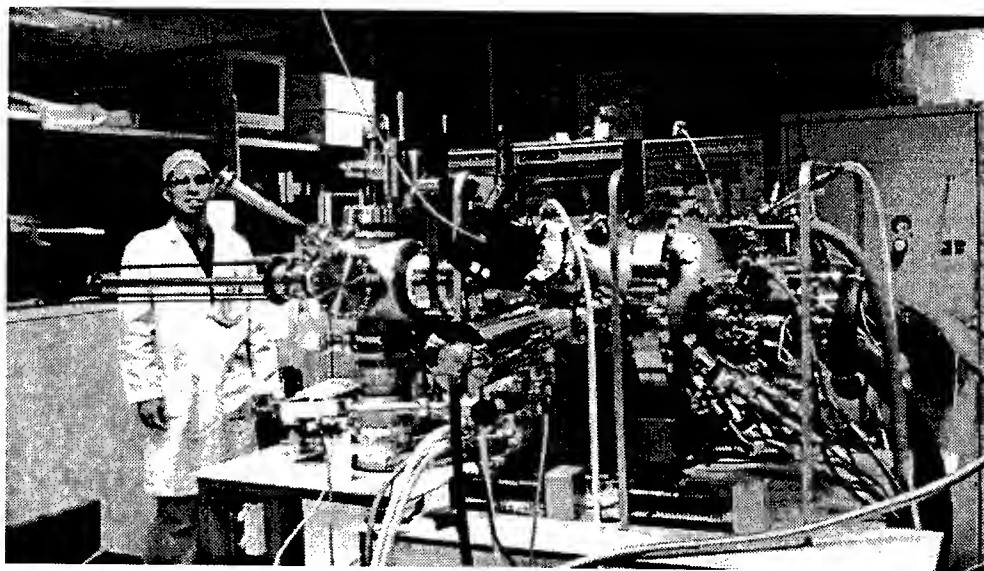


Fig. 4. Main growth chamber (on right) with all attachments. RGA (upper black box) and RHEED gun (just below) are shown in foreground.

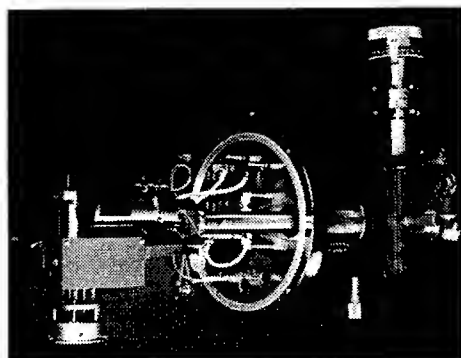
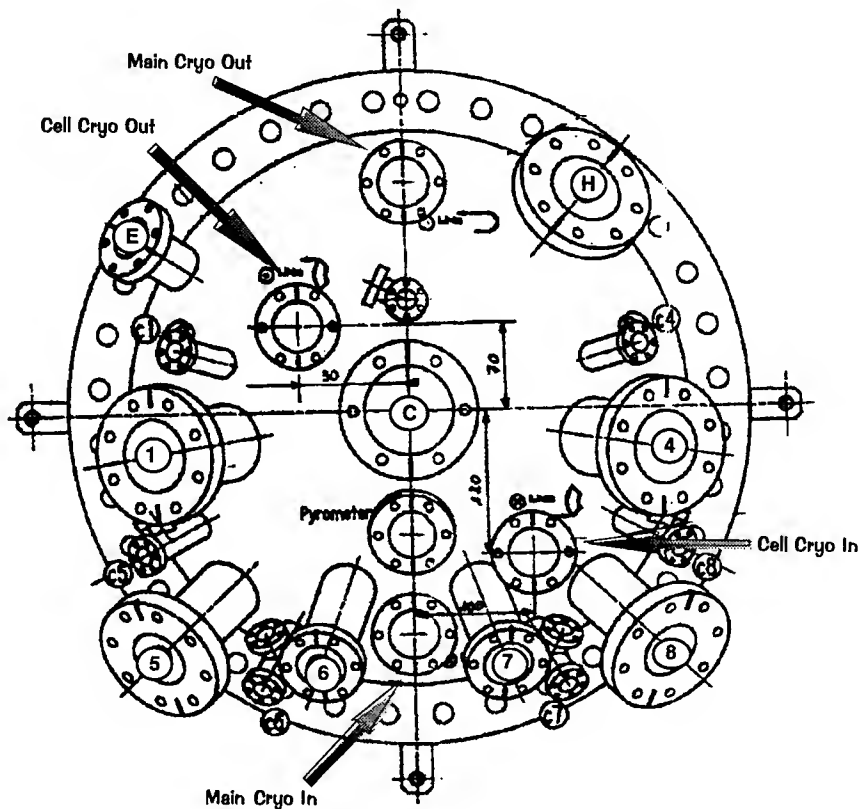


Fig. 5. Riber MBE-32 manipulator.

The cell panel (Fig. 6) contains both Knudsen effusion cells and a high temperature gas injector. In the near future, the cell panel will also be fitted with a plasma source to further to enable the growth of III-nitride wide band-gap semiconductors. Also on the cell panel is a viewport for a pyrometer (to be retrofitted), an XTC crystal thickness monitor, and shutter motors. The XTC is employed to measure film deposition rate from an electron beam gun used to melt source material. Analysis in the chamber is provided by an Inficon residual gas analyzer (RGA) and a Staib Instruments 35 kV reflection high energy electron diffraction (RHEED) gun. The RGA is used to determine the background contaminants in the chamber and the process species during growth. The RHEED gun is used to determine the crystallinity and growth rates on the substrate. This will be discussed in a later section.

## 2.4 SOURCES

Fig. 6 shows the setup of the cell panel and location of the sources on the MBE system. Ports 1,5-8 are fitted with Knudsen cells (shown schematically in Fig. 7) and to be filled with solid source ultra high purity (UHP) material for evaporation. An EPI high temperature cell resides in port 1 capable of temperatures near 2000°C and is scheduled to be filled with Er. Since port 1 is in a low angle position (5°), the crucible can only hold a small charge in comparison to the bottom row (5-8) at 32°. Ports 5 and 8 contain Riber single filament cells filled with Mg and Al, both dopants in the nitrides and SiC. Cells 6 and 7 are EPI Sumo dual filament cells designed to reduce cell "spitting" during operation



<u>PORTS</u>	<u>Flange/Clearance</u>	<u>Angle</u>	<u>Source</u>
[1]	4.5"/2.75"CFF—1.75"	5°	Er-EPI High Temp
[C]	4.5"CFF—2.5"	0°	N2 Plasma
[4]	4.5"/2.75"CFF—1.75"	5°	Gas Injector*
[5]	4.5"/2.75"CFF—1.75"	32°	Mg-Riber
[6]	2.75"CFF—1.75"	32°	Ga-EPI SUMO
[7]	2.75"CFF—1.75"	32°	In-EPI SUMO
[8]	4.5"/2.75"CFF—1.75"	32°	Al-Riber
[E]			Ellipsometer
[H]			XTC

Fig. 6. Schematic of growth chamber cell panel.

and increase uniformity. These will be filled with Ga and In for growth of nitrides and other semiconductor heterostructures. A three-line gas injector and cracker is installed in cell 4. The injector has 1200°C temperature capability, two Baratron flow controllers, and one mass flow controller. Each line has a manifold to switch between 2 different gases: line one has silacyclobutane (SCB  $\text{SiC}_3\text{H}_8$ ) and propane ( $\text{C}_3\text{H}_8$ ), line two has dilution  $\text{N}_2$  or ammonia ( $\text{NH}_3$ ), and line three has  $\text{H}_2$ . The center port on the chamber is reserved for a nitrogen plasma source for growing nitrides.

## 2.5 ELECTRONICS

A gas cabinet (3) to the left of the cell panel houses the cylinders for the three-line gas injector and a mechanical pump that evacuates the gas lines and the cryopump exhaust. The gas cabinet electronics, power supplies, pressure gauges, safety actuators, and temperature controllers are all located in racks (2) next to the system. Two control computers (1) are used to integrate the electronics: one PC operating in the NextStep Unix environment using a Riber program called Accessible to integrate all the growth components of the system; a second PC running under DOS/Windows to operate the RGA and RHEED gun.

## 2.6 ANALYSIS EQUIPMENT

In situ analysis is one crucial advantage that MBE has over other growth processes. This is made possible by the UHV conditions inside the growth chamber. The environment in the chamber can be monitored by the RGA before, during, and after the run. This can determine the quantities of reactive species in the chamber, help locate the origin of grown-in impurities, and evaluate the condition of the pumping. We have

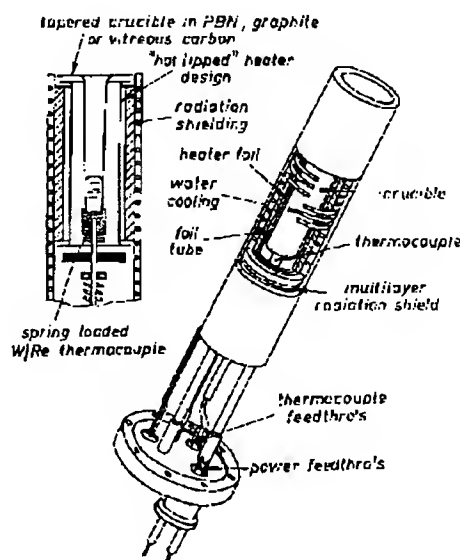


Fig. 7. Cut away view of standard effusion cell on MBE systems.

installed a Leybold Inficon H100M quadrupole with a Faraday cup for high pressure gas analysis and an electron multiplier for low ( $<10^{-6}$  Torr) pressures. For surface-sensitive analysis of the substrate, we have a Staib Instruments 35kV electron gun for obtaining RHEED patterns. This can quantify the substrate condition, from the effectiveness of the pre-cleaning and outgassing to the atomic growth characteristics and subsequent annealing effectiveness. RHEED can also be used to determine growth rates and surface reconstructions.

### 3. Epitaxy of SiC - Literature Survey

#### 3.1 Heteroepitaxy

Carbonization is an effective process to relax the large lattice mismatch between Si and SiC. There are two different carbonization processes. In the first process, the Si substrate is annealed to high temperature (750 - 1100 °C) and then hydrocarbon gas or carbon (atoms or ions) are introduced to react with Si surface. Several simple hydrocarbon molecules, such as  $C_2H_2$ ,  $C_2H_4$ ,  $C_3H_8$  have been commonly used as carbon source. The gas pressure during carbonization varied from  $10^{-9}$  to  $10^{-5}$  Torr. Some researchers found the carbonized films were polycrystalline by this process [3 - 5], while others obtained films of good quality [6 - 9].

In the second process, carbon containing gas is first introduced into the chamber until the desired pressure is reached, and then the substrate temperature is raised to high temperatures (750 - 1100 °C) at a slow ramp rate (5 - 25 °C/min). There are several versions of this process, with variations in the temperature ramp rate at different temperature stages and the incorporation of an oxide removal step [3]. In general, this process is very effective for obtaining single crystal 3C-SiC without double- positioning twin structures and pits [3, 4, 5]. This process probably seals off the outward Si diffusion from the Si substrate, which is believed to cause surface defects [4, 5].

Essential growth of SiC is used to further grow SiC on the carbonized film with both Si- and C- containing source. The substrate temperature ranges from 750 to 1100°C, the total pressure is usually  $10^{-7}$  to  $10^{-5}$  torr during growth. The effect of the flux ratio between Si and C species on film quality and stoichiometric ratio has been extensively studied [5,10-13]. It was found that  $J_{Si}/J_C > 1$  is a general rule for growing good crystalline

3C-SiC film with 1:1 stoichiometric ratio. Atomic layer epitaxy by MBE has also attracted a lot of attention [14-17]. It was found that the surface superstructures during an alternating supply of C source and Si source can be used to control film growth to atomic level accuracy.

### **Lattice matched growth**

The lattice constant of (111) 3C-SiC nearly matches ( $< 0.1\%$ ) that of the c-plane of 6H-SiC. 6H-SiC is commercially available from several sources, with Cree Research being the main supplier. 6H-SiC is also used as a substrate for growing good quality SiC films. Some results show that the epitaxially grown 3C-SiC (111) on 6H-SiC (1000) at 850 -1000 °C have double-positioning twin structure [18- 19], while on 6H-Si (0114) substrate, the 3C-SiC(100) epilayers were grown without twin structures at temperatures as low as 850°C [19].

Homoepitaxy of 6H-SiC was also investigated and growth process controlled to an atomic level was obtained by monitoring surface superstructures during the supply of Si and C atoms [20-22]. The grown film is predominantly 6H-SiC with small amount of 3C-SiC mainly located at defect sites. These defects mostly extend from the substrate into the film.

## **4. PRELIMINARY RESULTS**

In this section we describe the first SiC MBE experiments performed at the University of Cincinnati and discuss our preliminary results.

### **4.1 EXPERIMENTAL CONDITIONS**

The epitaxial growth of 3C-SiC was carried in a Riber GSMBE 32 system. The detailed description of the system was described above. Briefly, the epitaxial growth process starts with the growth chamber being evacuated by the Ti sublimation pump and cryogenic pump. After 60 hours baking at 250 °C, a base pressure of  $3 \times 10^{-11}$  torr can be routinely obtained without filling the cryoshroud with liquid nitrogen. A quadrupole mass

analyzer and a 35 keV reflection high-energy electron diffraction (RHEED) system are mounted on the growth chamber for gas analysis and surface crystallinity characterization, respectively. A high temperature gas injector (1200 °C) with three gas lines is employed to introduce propane and silacyclobutane ( $\text{SiC}_3\text{H}_8$  - SCB) into the growth chamber. The gas flow is controlled by a PID controller. The substrate is introduced through the load lock chamber which was pumped by a Ti sublimation pump and an ion pump to  $< 5 \times 10^{-10}$  Torr. The initial outgassing of the sample is also performed in this chamber. The sample was transported by a magnetically coupled transfer rod into the growth chamber and locked onto a five-dimensional manipulator which contains a high temperature (1200 °C), high uniformity oven. The heater temperature is measured by a W-Re thermocouple and controlled by a PID controller.

Two preliminary growth experiments have been carried out. The growth conditions are summarized in Table 1. In the first experiment a carbonized SiC SOI (Si On Insulator) wafer with (111) orientation was used as the substrate for SiC MBE growth. The carbonization was first performed by rapid thermal CVD. The carbonized SOI sample was cleaned by dipping into 1% HF for 1 min and rinsing in DI water for 2 min. The sample was then introduced into the load lock chamber, where it was outgassed at 300 °C for 12 hours. After the sample was transferred into growth chamber it was heated to 1000 °C until a sharp RHEED pattern with 6-fold rotation symmetry was observed, indicating a clean  $\beta$ -SiC surface. The further growth on the carbonized film was done by SCB pyrolysis with a sample temperature of 1000 °C.

#	Substrate	Carbonization by $\text{C}_3\text{H}_8$				SCB Growth by MBE		
		type	time (min)	pressure (torr)	temp (°C)	time (min)	pressure (torr)	temp (°C)
C21	Si(111) (SOI)	CVD		760	1235	80	$\sim 3 \times 10^{-7}$	1000
C1	Si(100) on-axis	MBE	120	$1.7-4.8 \times 10^{-6}$	776			
CE1	Si(100) on-axis	MBE	20	$1.3-7.1 \times 10^{-6}$	800	9	$1.8-3.7 \times 10^{-6}$	800
CE2	Si(100) on-axis	MBE	120	$1.4-5.1 \times 10^{-6}$	800	95	$1.8-3.5 \times 10^{-6}$	800

Table 1. Experimental conditions for SiC MBE growth.

The second experiment utilized a Si(100) wafer as substrate. Both carbonization and essential growth were carried out in the MBE system. Propane was used for carbonization and SCB for essential growth on the carbonized film. The temperature program is shown in Fig. 8.

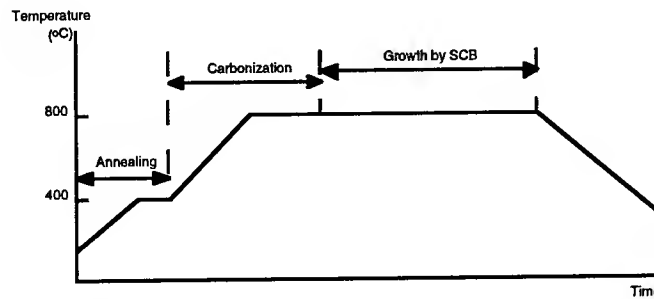


Fig. 8. Temperature program of SiC growth process by MBE.

In this temperature program diagram, there are three parts: (1) the Si(100) substrate was heated to 400 °C in UHV ( $10^{-10}$  -  $10^{-9}$  torr) with a ramp rate of 20 °C/min; (2) the supply of the propane was started and then the substrate temperature was ramped up to 800 °C at a rate of 20 °C/min; (3) the essential growth of 3C-SiC on the carbonized layer was done at 800 °C with SCB for both carbon and silicon source.

The crystallographic features of the growing surface were monitored by the RHEED system. The RHEED images were captured by a CCD camera system, and a sophisticated software (RHEED-VISION) was used to grab the image and monitor up to four diffraction spot intensities during the growth. The thickness and composition of the grown layers were measured by secondary ion mass spectrometry (SIMS) using  $\text{Cs}^+$  bombardment with positive ion detection. Elements monitored were C, Si, O, N and B. Relative sensitivity factors derived from a SiC standard were used to convert ion counts to concentrations. The thickness of SiC layer was estimated by finding the position in the carbon depth profile at which atomic carbon concentration is 50% of its maximum. The SCB growth rate was calculated by:

$$\text{Growth Rate} = \frac{\text{Thickness}(\text{total}) - \text{Thickness}(\text{carbonized})}{\text{Time}(\text{SCB})} \quad (1)$$



Propane was supplied by Matheson Gas Products, Inc. with a purity of 99.97%. SCB was provided by Dow Corning. SCB is a liquid at room temperature with a vapor pressure of 400 Torr. Both propane and SCB vapor were used without further purification. Before gas or vapor were introduced into the growth chamber, the cryoshroud was filled with liquid nitrogen, the Ti-sublimation pump and the RGA were turned off to obtain a cleaner growth environment, all cells and gas injector were set to 200 °C to avoid their contamination by C<sub>3</sub>H<sub>8</sub> or SCB. Both propane and SCB were introduced towards the substrate through the gas injector. The flow rate of propane during carbonization was equivalent to its partial pressure from 1.5 - 5.5 x10<sup>-6</sup> torr, the flow rate of SCB during growth corresponded to its partial pressure from 6.1 to 16 x10<sup>-7</sup> torr. The sample was not rotated during the growth in order to be able to monitor the RHEED pattern.

#### 4.2 RESULTS AND DISCUSSION

Fig. 9 shows a typical RHEED pattern of the 3C-SiC sample (C21) obtained by RTCVD carbonization. This pattern has twin diffraction spots which may be the superposition of two diffraction patterns corresponding to two domains which are 180° rotated from each other. The schematic of this interpretation is shown in Fig. 10.

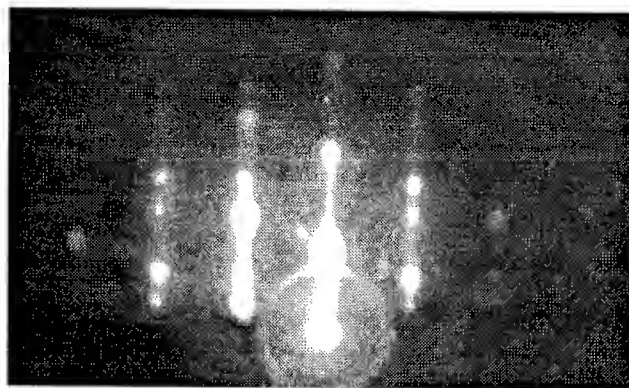


Fig. 9. Typical RHEED pattern of 3C-SiC film by carbonization of SOI with RTCVD.

Therefore, the RTCVD carbonized film appears to have double-positioning twin structure. The possible reason of this twin structure is that the Si carbonization is driven by reaction on terraces, with different terraces causing the difference in stacking order.

After the carbonization by CVD as shown in Fig. 9, essential growth of SiC was performed in MBE growth chamber with SCB as both Si and C source. During the growth, some fluctuation of intensity was observed. No regular intensity oscillation was observed, probably indicating that the growth was not in a layer-by-layer mode or the growth rate is too small. However, the 3C-SiC RHEED pattern was very clear during the entire growth period. Fig. 11 shows a RHEED image after 80 min growth at a SCB partial pressure of  $3.0 \times 10^{-7}$  Torr. Compared with the RHEED pattern of Fig. 9, the twin spots in Fig. 11 are much weaker, which probably indicates the double-positioning twin structure is much reduced after essential growth by MBE.

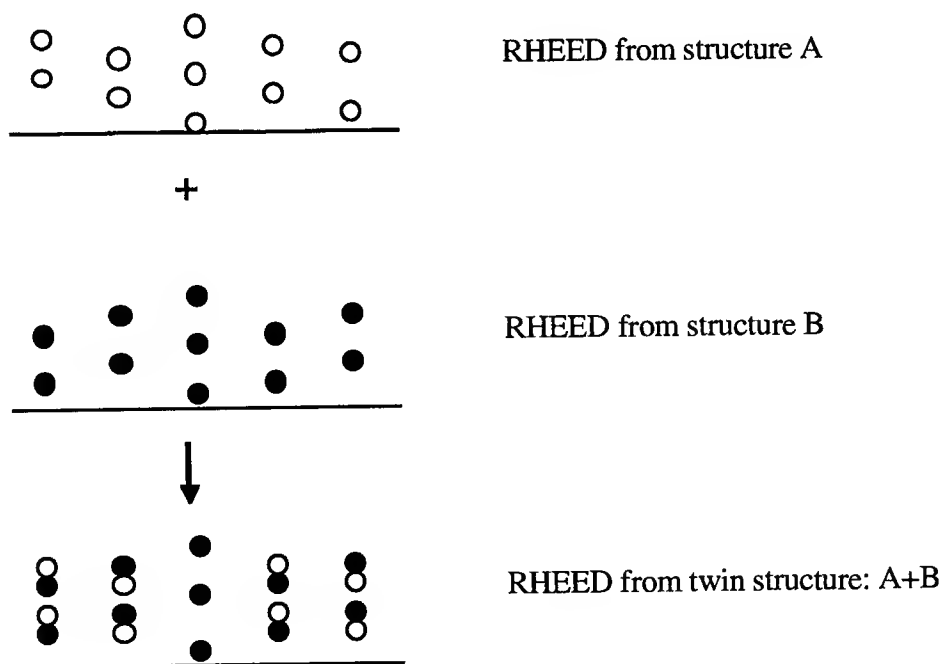


Fig. 10. Interpretation of the RHEED pattern in Fig. 9.

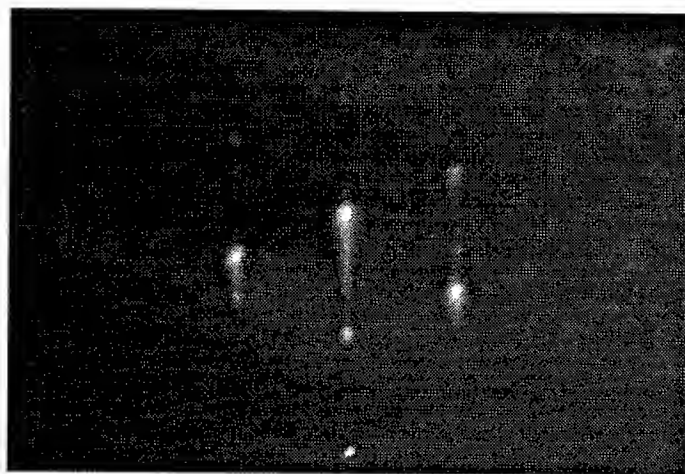


Fig. 11. RHEED pattern of SiC film grown by MBE with SCB on SOI sample carbonized with propane by CVD (C21).

Carbonization and essential growth were also performed sequentially in MBE system. The substrate was a 2 inch on-axis Si(100) wafer. Fig. 12 shows the RHEED patterns observed during SiC growth of sample CE2. Fig. 12 (a) is the RHEED pattern of the Si(100) surface after annealing at 400 °C for half a hour. The clear streaky pattern indicates a clean Si(100) surface. After introducing propane into the growth chamber, the temperature of substrate was ramped at a rate of 20 °C/min. During this temperature ramping process, the RHEED pattern did not change until 760 °C. As the carbonization proceeded, the RHEED pattern associated with 3C-SiC became stronger while that of Si(100) became weaker and finally disappeared. Fig. 12 (b) is the 3C-SiC RHEED pattern obtained after 2 hours carbonization of Si(100). No twin spots were observed in the MBE carbonized film as were seen in the carbonized film by RTCVD. This observation indicates that carbonized film by MBE has fewer double-boundary defects than that obtained by CVD. The further growth of 3C-SiC was done by introducing SCB at a sample temperature of 800 °C. Fig. 12 (c) shows the RHEED pattern after 5 min growth by SCB. The pattern looks more streaky than that of carbonized film and indicates the surface is smoother. As growth proceeded, the rings started to appear as shown in Fig. 12 (d), which indicates a poly-crystal SiC film was being formed. The possible reason is that the

strain due to the large lattice mismatch between Si and SiC was not completely released by carbonization.

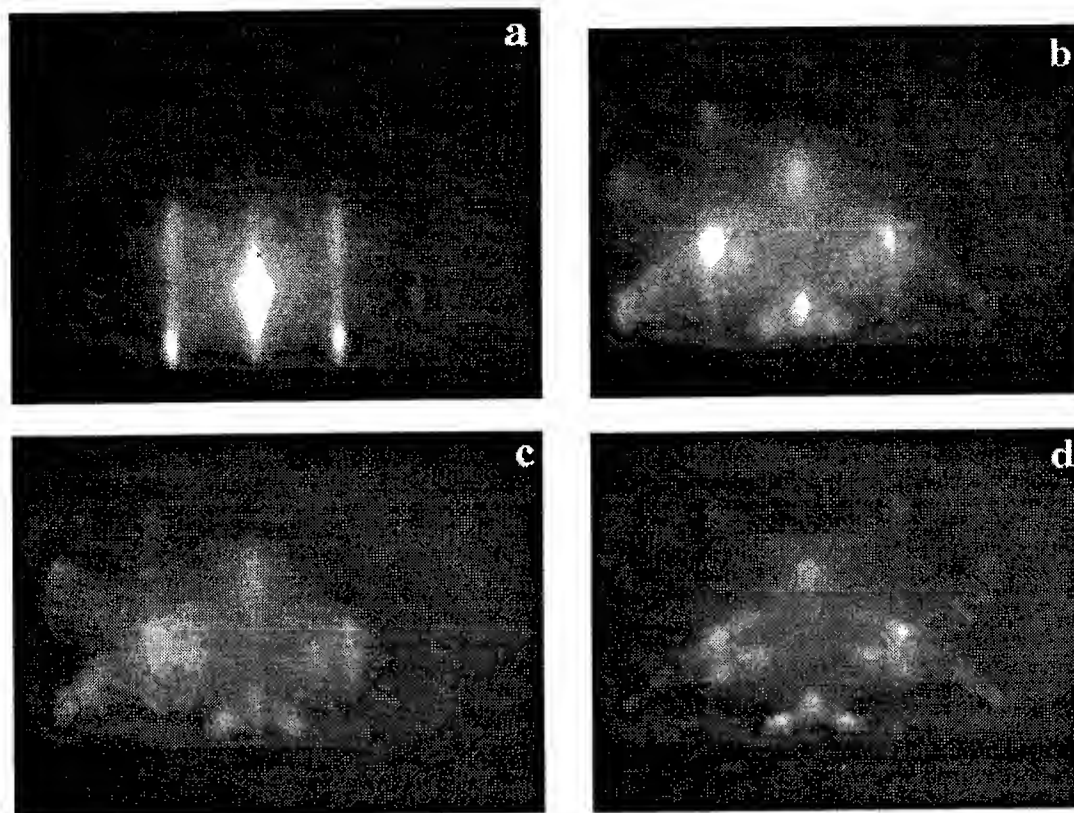


Fig. 12. RHEED patterns observed during SiC growth (CE2): (a) clean Si(100) surface at 400°C; (b) 3C-SiC grown by  $C_3H_8$  carbonization for two hours at  $P_{C_3H_8} = \sim 3 \times 10^{-6}$  torr; (c) 3C-SiC grown on the carbonized layer by SCB for 5 min at  $P_{SCB} = 1.4 - 2.9 \times 10^{-7}$  torr; (d) 3C-SiC grown on the carbonized layer by SCB for 95 min at  $P_{SCB} = 1.8 - 3.5 \times 10^{-6}$  torr.

The film composition and thickness of the MBE-grown SiC layers were measured with secondary ion mass spectrometry (SIMS). Fig. 13 shows the depth profile of the 3C-SiC film (sample C1) grown by MBE carbonization of Si(100) at 776°C for two hours. It is apparent that only a very thin Si layer (probably  $\sim 30-40$  Å) has been converted at this temperature. It is interesting to point out, however, that even this very thin carbonized layer produced a very good (similar to Fig. 12b) RHEED pattern indicative of crystalline

SiC. Fig. 14 contains the depth profile of the 3C-SiC film (CE1) grown at 800°C using carbonization for 20 min followed by SCB growth for 9 min. The thickness of the SiC film is approximately 95-100 Å judging by the depth at which the carbon concentration reaches the 50% point. Fig. 15 shows the depth profile of the SiC film (CE2) also grown at 800°C, but for significantly longer times: carbonization for 120 min and essential growth by SCB for 100 min. In this sample we find a SiC film thickness of around 600 Å. Using the film thickness obtained from the SIMS depth profiles and the SCB growth time, the SiC growth rate by SCB at 800 °C is estimated to be around 0.1 Å/s. This low growth rate could be due to a combination of effects: low growth temperature, low growth pressure, and possibly the fluctuation of the gas flow during gas introduction. In Figs. 14 and 15, a long carbon tail into the Si substrate is observed. This could be caused by several effects: (a) absence of an abrupt interface between SiC and Si substrate, due to carbon atoms diffusing into the Si during growth; (b) non-uniform film thickness resulting in certain locations where the SiC layer is removed sooner during SIMS profiling. The SIMS data also show that impurities N, O and B concentration are less than 0.1% in the SiC film.

## 5. SUMMARY AND FUTURE WORK

In this report we described the Riber MBE 32 system recently installed in the Nanoelectronics Laboratory at University of Cincinnati. Some preliminary results of heteroepitaxy of SiC with this MBE system were presented. We have clearly demonstrated that SiC growth with SCB can be accomplished by the MBE technique. At 800 °C, the SiC growth rate obtained with SCB under the first set of conditions utilized is quite low (0.1 Å/s). We are continuing with experiments designed to determine the effect of growth temperature, pressure and flow rate on the growth process: growth rate and film quality of SiC. We plan to also investigate the use of other gas precursors for SiC growth and the in-situ sequential growth of SiC and GaN. In addition to crystallinity and surface morphology, we will also investigate the electrical and optical properties of the MBE-grown SiC and GaN films.

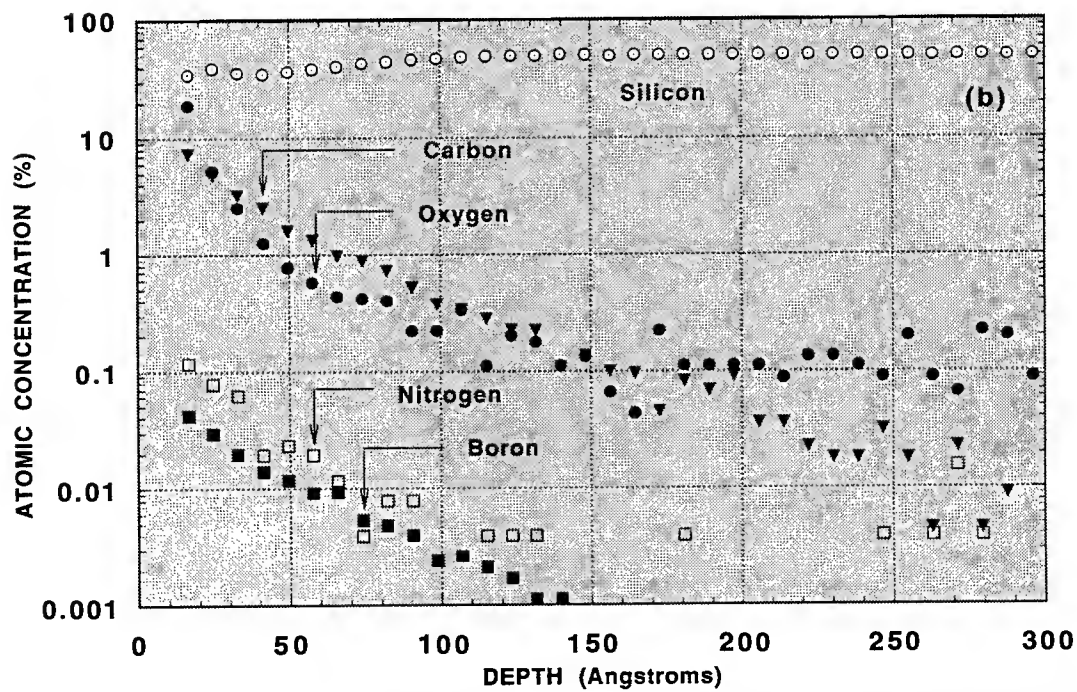
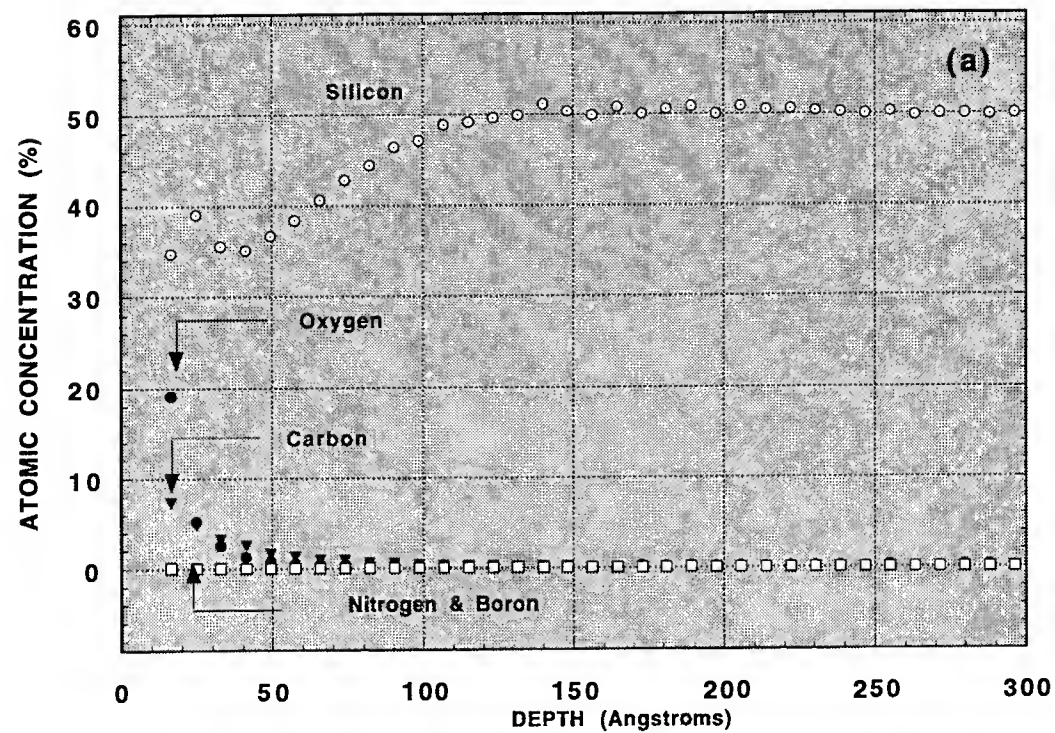


Fig. 13. SIMS depth profile of SiC film (sample C1) grown by two hours carbonization with  $C_3H_8$ . (a) linear scale; (b) semi-log scale.



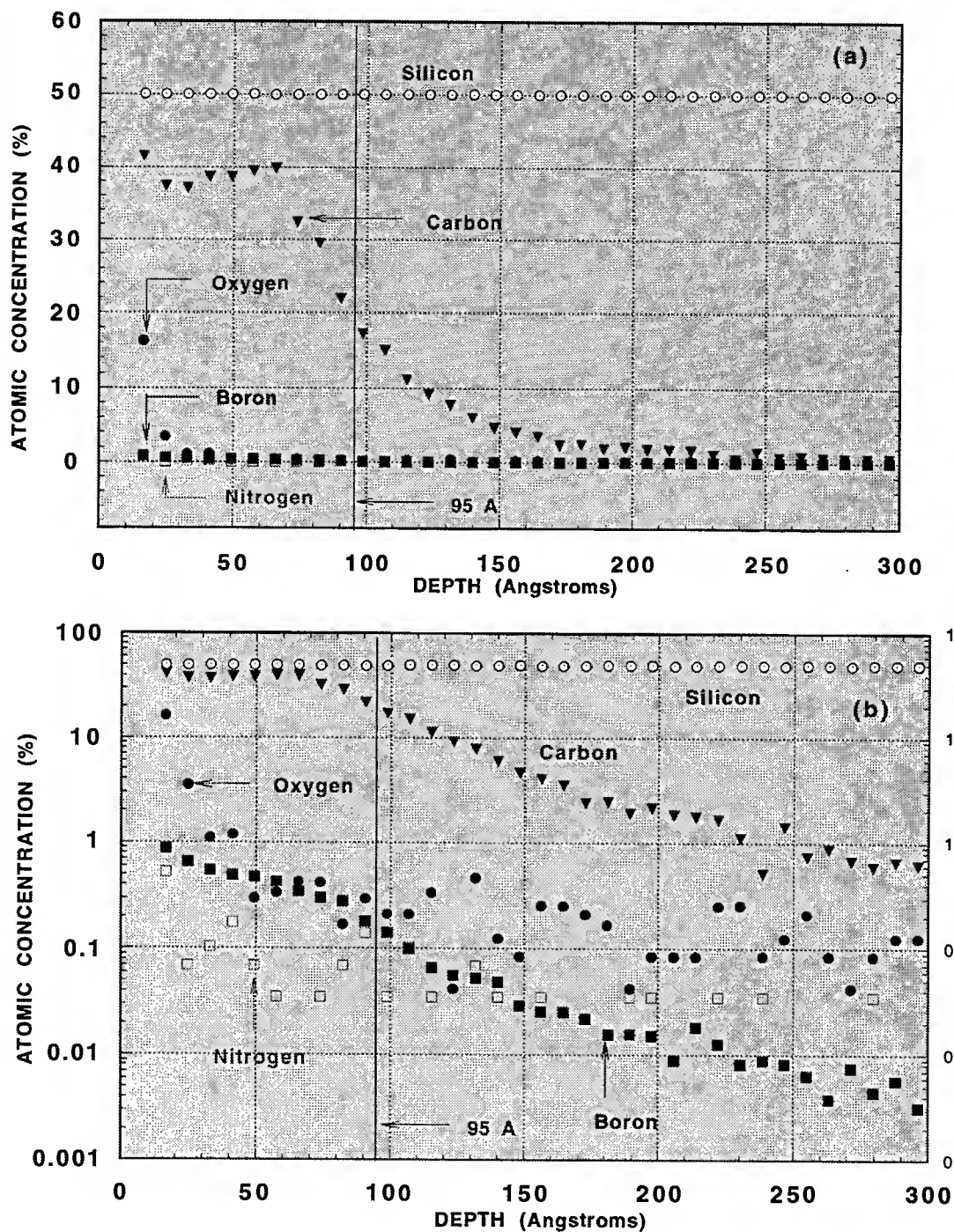


Fig. 14. SIMS depth profile of SiC film (sample CE1) grown by 20 min carbonization and 9 min essential growth by SCB. The SiC film thickness is 95 Å. (a) linear scale; (b) semi-log scale.

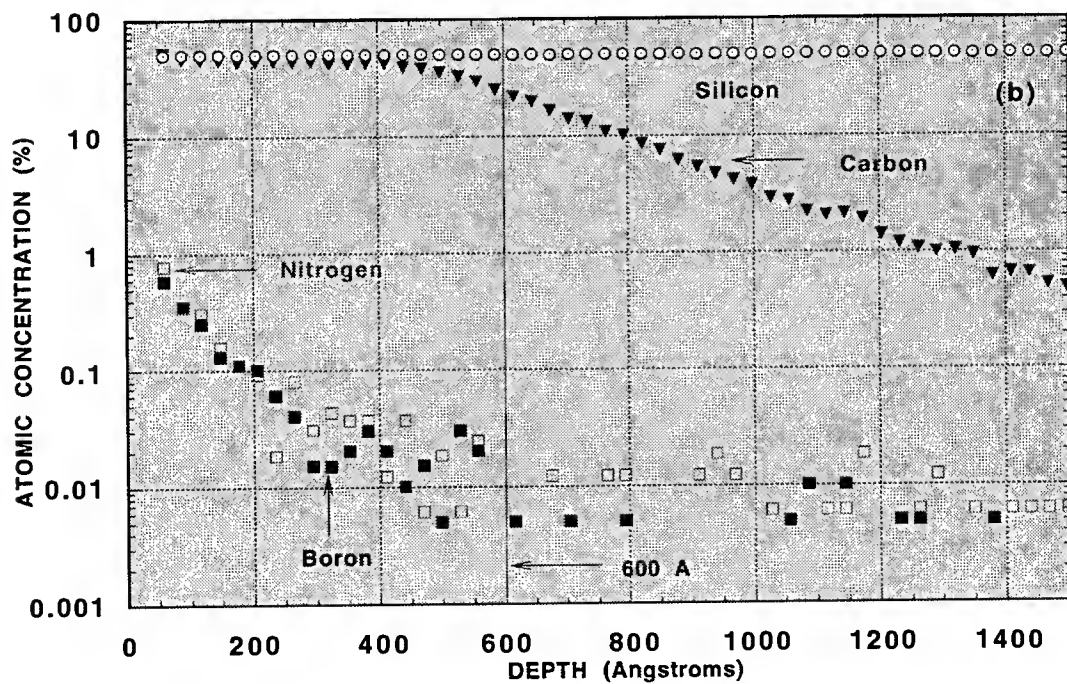
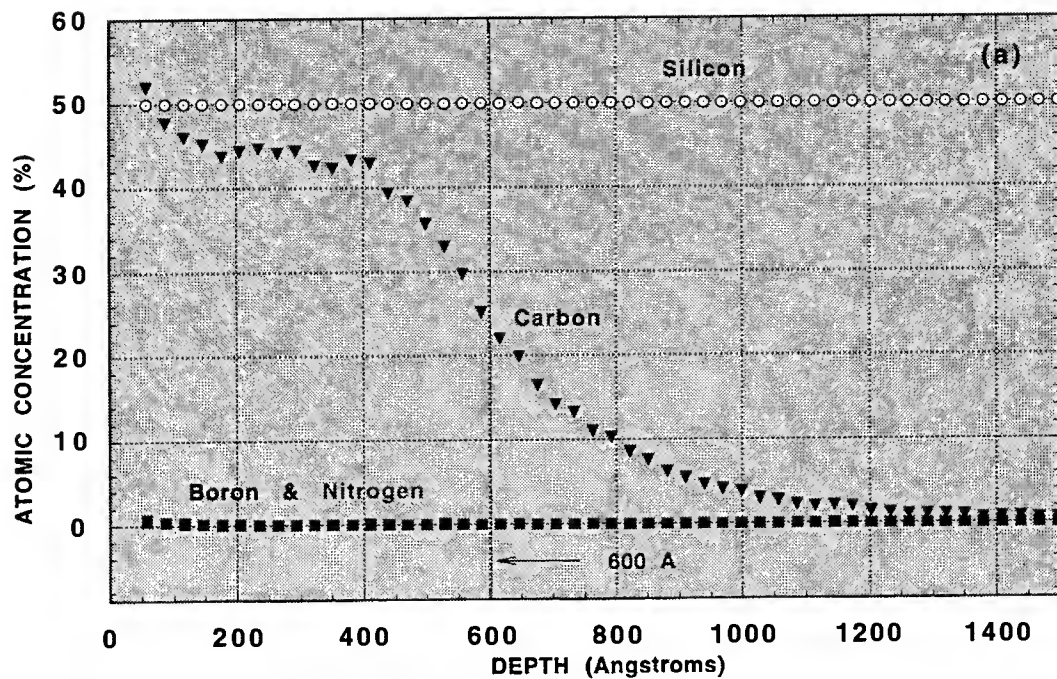


Fig. 15. SIMS depth profile of SiC film (# CE2) grown by 120 min carbonization and 100 min essential grown by SCB. The film thickness is 600 Å. (a) linear scale; (b) semi-log scale.



## 6. REFERENCES

1. W. E. Knippenberg, Philips Res. Rep., **18** (1963) 161.
2. W. E. Nelson, F. A. Halden and A. Rosengreen, J. Appl. Phys., **37** (1966) 333.
3. T. Yoshinobu, T. Fuyuki and H. Matsunami Jpn. J. Appl. Phys, **30** (1991) L1086.
4. T. Yoshinobu, H. Mitsui, Y. Tarui, T. Fuyuki and H. Matsunami, J. Appl. Phys. **72** (1992) 2006.
5. S.-I. Motoyama, N. Morikawa, M. Nasu and S. Kaneda, J. Appl. Phys., **68** (1990) 101.
6. G.L. Zhou, Z. Ma, M. E. Lin, T. C. Shen, L. H. Allen and H. Morkoc J. Crystal Growth, **134** (1993) 167.
7. T. Sugii, T. Aoyama and T. Ito J. Electrochem. Soc., **137** (1990) 989.
8. K. Kim, Si-Choi and K.L. Wang, J. Vacuum Sci. Technol. B **10** (1992) 930.
9. K. Kim, Si-Choi and K.L. Wang, Thin Solid films, **225** (1993) 235.
10. S.-I. Motoyama, N. Morikawa and S. Kaneda, J. Cryst. Growth **100** (1990) 615.
11. S. Tanaka, R. S. Scott and R.F. Davis, Appl. Phys. Lett. **65** (1994) 2851.
12. S. Motoyama and S. Kaneda Appl. Phys. Lett. **54** (1989) 242.
13. S. Kaneda, Y. Sakamoto, C. Nishi, M. Kanaya and S. Hanai, Jpn. J. Appl. Phys. **25** (1986) 1307.
14. T. Fuyuki, T. Yoshinobu and H. Matsunami, Thin Solid Films, **225** (1993) 225.
15. H. Matsunami, Physica B, **185** (1993) 65.
16. T. Yoshinobu, M. Nakayama, H. Shiomi, T. Fuyuki and H. Matsunami, J. Cryst. Growth, **99** (1990) 520.
17. T. Fuyuki, M. Nakayama, T. Yoshinobu, H. Shiomi and H. Matsunami, J. Cryst. Growth, **95** (1989) 461.
18. L. B. Rowland, R.S. Kern, S. Tanaka and R.F. Davis, J. Mater. Res. **8** (1993) 2753.
19. T. Yoshinobu, H. Mitsui, I. Izumikawa, T. Fuyuki and H. Matsunami, Appl. Phys. Lett., **60** (1992) 824.
20. A. Fissel, B. Schroter and W. Richter Appl. Phys. Lett. **66** (1995) 3182.
21. A. Fissel, U. Kaiser, E. Ducke, B. Schroter and W. Richter, J. Crystal Growth, **154** (1995) 72.
22. A. Fissel, U. Kaiser, K. Pfennighaus, B. Schroter and W. Richter Appl. Phys. Lett. **68** (1996) 1.

**PERFORMANCE OF ITERATIVE AND NONITERATIVE SCHEMES  
FOR IMAGE RESTORATION AND SUPERRESOLUTION PROCESSING  
IN MULTISPECTRAL SEEKER ENVIRONMENTS**

**Malur K. Sundareshan**

Professor of Electrical and Computer Engineering  
University of Arizona  
Tucson, AZ 85721

Final Report for:  
AFOSR Summer Research Extension Program  
Wright Laboratory Armament Directorate

Sponsored by:  
Air Force Office of Scientific Research  
Bolling Air Force Base, DC

and

Wright Laboratory Armament Directorate  
Eglin Air Force Base, FL

February 1997

**PERFORMANCE OF ITERATIVE AND NONITERATIVE SCHEMES  
FOR IMAGE RESTORATION AND SUPERRESOLUTION PROCESSING  
IN MULTISPECTRAL SEEKER ENVIRONMENTS**

**Malur K. Sundareshan**

Professor of Electrical and Computer Engineering

University of Arizona

Tucson, AZ 85721

**ABSTRACT**

Restoration and superresolution processing of images collected from various sensors deployed in multispectral seeker environments often become necessary to enhance image resolution in order to facilitate better false target rejection, improved automatic target recognition and aimpoint selection. Due to the critical importance of this technology to several Air Force missions and its relevance to diverse on-going projects in several Air Force laboratories, we are conducting detailed studies on this topic. Two major outcomes from these studies have been the recognition of the importance of optimally tailored restoration and superresolution algorithms for the individual sensor type and operating conditions, and the feasibility of using iterative and noniterative processing techniques in an intelligent tailoring of these algorithms. Use of these techniques however can result in different performance levels and also can bring specific advantages and disadvantages to the overall restoration and superresolution function (and hence to the surveillance and smart munition guidance objectives). The principal objective of the research reported here is to give a qualitative comparison of the performance expected from an iterative procedure based on Bayesian estimation methods with that resulting from a class of noniterative restoration methods. More quantitative performance evaluations directed to an explicit demonstration of the spectrum extrapolation in both one-dimensional and two-dimensional signals from an iterative implementation of a Maximum Likelihood (ML) algorithm are also presented. A modification of the algorithm to facilitate simultaneous estimation of point spread function of sensor and resolution enhancement of input image is outlined and an illustration of its performance in processing a set of passive millimeter-wave (MMW) images obtained from a 95 GHz 1-foot diameter aperture radiometer is given.

**PERFORMANCE OF ITERATIVE AND NONITERATIVE SCHEMES  
FOR IMAGE RESTORATION AND SUPERRESOLUTION PROCESSING  
IN MULTISPECTRAL SEEKER ENVIRONMENTS**

**Malur K. Sundareshan**

**1. INTRODUCTION**

A significant problem that affects the successful realization of the goals of many tactical missions is the poor resolution of images collected from the sensors used to assist surveillance and guidance operations. The problem is particularly prevalent in autonomous missile guidance applications where diverse mission requirements, such as reliable target detection, classification, interleaved acquisition, track and engage modes, and precision kill, critically depend on the quality of data collected from the sensors deployed in missile seekers. While it is true that deployment within a common aperture package of complementary sensors operating over different frequency ranges would enhance the detection, classification and track maintenance performance of missile seekers (in addition to providing increased fault tolerance and greater immunity to countermeasures), such multispectral environments also accentuate other considerations, such as sensor fusion requirements, which in turn demand high resolution sensor data.

The problem of poor resolution in imaging sensors stems mainly from deployable antenna size limitations (which preclude simply increasing the physical aperture of sensors to gain high image resolution) and the consequent diffraction limits on the achievable resolution. It may be noted that the wavelength of a synthetic aperture radar (SAR) operating at 1GHz is about 1 inch long and one needs an antenna as big as 40 ft wide in order to achieve a resolution requirement of being able to distinguish points in a scene separated by about 1 meter at a distance of 1 Km [1]. Passive millimeter-wave (PMMW) sensing offers superior adverse weather capabilities (over infra-red (IR) sensors, for instance) due to easy penetration through fog, dust, smoke, etc. However, PMMW image acquisition sensors suffer from poor angular resolution. It is well documented that the angular resolution achievable by a 94 GHz system with a 1 ft diameter antenna is only about 10 mrad, which translates into a spatial resolution of about 10 meters at a distance of 1 Km. Some recent studies [2] have also established that for ensuring reasonably adequate angular resolution (typically of the order of 4 mrad), a 94 GHz PMMW imaging system with a sensor depression angle of  $60^\circ$  -  $80^\circ$  needs to be confined to very low operational altitudes (of the order of 75-100 meters) which puts inordinate demands on the guidance schemes to facilitate such requirements. Similar resolution limitations and the consequent requirements on operational conditions (some of which may be clearly impossible to satisfy for tactical missions with reliability and survivability constraints) exist for the other types of sensing modalities as well.

Typical seeker antenna patterns are of a "low-pass" filtering nature due to the finite size of the antenna or lens that makes up the imaging system and the consequent imposition of the underlying diffraction limits. Hence the image recorded at the output of the imaging system is a low-pass filtered version of the original scene. The portions of the scene that are lost by the imaging system are the fine details (high frequency spectral components) that accurately describe the objects in the scene, which also are critical for reliable detection and classification of targets of interest in the scene. Hence some form of image processing to restore the details and improve the resolution of the image will invariably be needed. Traditional image restoration procedures (based on deconvolution and inverse filtering approaches) attempt mainly at reconstruction of the passband and possibly elimination of effects of additive noise components. These hence have only limited resolution enhancement capabilities. Greater resolution improvements can only be achieved through a class of more sophisticated algorithms, called superresolution algorithms, which provide not only passband resolution but also some degree of **spectral extrapolation**, thus enabling to restore the high frequency spatial amplitude variations relating to the spatial resolution of the sensor and lost through the filtering effects of the seeker antenna pattern. A tactful utilization of the imaging instrument's characteristics and any *a priori* knowledge of the features of the target together with an appropriately crafted nonlinear processing scheme is what gives the capability to these algorithms for superresolving the input image by extrapolating beyond the passband range and thus extending the image bandwidth beyond the diffraction limit of the imaging sensor.

For application in missile seeker environments, it must be emphasized that superresolution is a post-processing operation applied to the acquired imagery and consequently is much less expensive compared to improving the imaging system for desired resolution. As an example, it may be noted that for visual imagery acquired from space-borne platforms, some studies indicate that the cost of camera payload increases as the inverse 2.5 power of the resolution. Hence a possible two-fold improvement in resolution by superresolution processing in this application roughly translates into a reduction in the cost of the sensor by more than 5 times. Similar relations also exist for sensors operating in the other spectral ranges (due to the relation between resolution and antenna size), confirming the cost effectiveness of employing superresolution algorithms. The principal goal of superresolution processing in multispectral seekers is hence to obtain an image of a target of interest (such as a mobile missile-launcher or a tank) via post-processing that is equivalent to one acquired through a more expensive larger aperture sensor.

Most of the recent analytical work in the development of image restoration and superresolution algorithms has been motivated by applications in Radioastronomy and Medical Imaging. While this work has given rise to some mathematically elegant approaches and powerful algorithms, a certain degree of care should be exercised in adapting these approaches and algorithms to the missile seeker environment. This is due to the convergence problems often encountered by iterative schemes and the specific statistical models representing the scenarios facilitating their development. For example, a slowly converging algorithm that ultimately guarantees the best resolution in the processed image may pose no implementational problems in Radioastronomy; however, it could be entirely unrealistic for implementation in an autonomous unmanned tactical system that must operate fast

enough to track target motion. Hence a careful tailoring of the processing algorithm for each sensor supporting the multispectral seeker is of critical importance in order to realize the possible performance benefits from superresolution processing which include better false target rejection, improved automatic target recognition and aimpoint selection.

The high degree of importance this topic has to present and future Air Force missions is clearly evident. Equally evident is the fact that research on this topic has an immediate application to a number of on-going programs in various Air Force laboratories. The research described in this report is an extension of the investigations that were conducted under a summer faculty visit to the Wright laboratory Armament Directorate. Two principal outcomes from these investigations [3] have been the recognition of the importance of optimally tailored restoration and superresolution algorithms for the individual sensor type and operating conditions, and the feasibility of using iterative and noniterative processing techniques in the tailoring of these algorithms. Use of these techniques however, can result in different performance levels and further entail specific advantages and disadvantages. In this report we shall give a qualitative comparison of the performance expected from an iterative procedure based on Bayesian estimation methods with that resulting from a class of noniterative restoration methods. Quantitative performance evaluations directed to demonstrating the spectrum extrapolation in both one-dimensional and two-dimensional signals resulting from an iterative implementation of a Maximum Likelihood (ML) algorithm will also be presented. To counter the inaccuracies in the modeling of the Point Spread Function (PSF) of the sensor, a modification of this algorithm to facilitate simultaneous estimation of PSF parameters and resolution enhancement of input image is outlined and an illustration of its performance in processing a set of passive millimeter-wave (MMW) images obtained from a 95 GHz 1-foot diameter aperture radiometer is given.

## **2. MATHEMATICAL REPRESENTATION OF IMAGE RESTORATION AND SUPERRESOLUTION PROBLEMS**

In this section we shall briefly describe the technical problems underlying the restoration and superresolution processing of sensor data in terms of reconstructing the spectral components of the signals being processed. A brief outline of the information available for developing specific algorithms in order to solve these problems will also be given.

### **2.1 Image Formation Process (Observation Model)**

Every systematic image processing study (including image restoration and superresolution processing) will start with an appropriate mathematical model characterizing the process of image formation by the sensor employed, which is termed an "observation model". Irrespective of the type of sensor actually used, a commonly used observation model takes the form

$$g = s(Hf) + n \quad (1)$$

where  $f$  denotes the object being sensed,  $g$  its image and  $H$  denotes the operator that models the filtering process including any associated degradations (such as due to small aperture size of the sensor) and blur phenomena (caused by atmospheric effects, motion of the object or the sensor, or out of focus operations, etc.).  $n$  denotes the additive random noise in the sensing process, which includes both the receiver noise and any quantization noise. The response of the image recording sensor to the intensity of input signal (light, radar, etc.) is represented by the memoryless mapping  $s(\cdot)$ , which is in general nonlinear.

For the sake of precision, let us consider the image to be obtained from an incoherent sensor. We will also assume that the image to be processed consists of  $M \times M$  equally spaced grey level pixels, obtained through a sampling of the image field at a rate that satisfies the Nyquist criterion. Furthermore, for mathematical tractability we will make the commonly used assumptions, which include: (i) space-invariant imaging process, (ii) ignore the nonlinear effects of the sensor, and (iii) approximate the noise process by a zero-mean white Gaussian random field which is independent of the object. With these assumptions, Equation (1) can be rewritten to relate the image intensity value  $g(i, j)$  at pixel  $(i, j)$  to the object pixel values as

$$g(i, j) = \sum_{(k, l) \in S} h(i - k, j - l) f(k, l) + n(i, j) \quad i, j = 1, 2, \dots, M \quad (2)$$

where  $h(i, j)$  denotes the point spread function (PSF) of the sensor.

For an image of size  $M \times M$ , Equation (2) corresponds to a set of  $M^2$  scalar equations specifying the formation of each image pixel. For a further simplified representation [4,5], by a lexicographical ordering of the signals  $g, f$  and  $n$ , one can rewrite Equation (2) as resulting from a convolution of two one-dimensional vectors

$h = [h(1), h(2), \dots, h(N)]^T$  and  $f = [f(1), f(2), \dots, f(N)]^T$  as

$$g(i) = h(i) \otimes f(i) + n(i) = \sum_{j=1}^N h(i - j) f(j) + n(i), i = 1, 2, \dots, N. \quad (3)$$

where  $N = M^2$ . More compactly, Equation (3) can be rewritten as the vector equation

$$g = Hf + n \quad (4)$$

where  $g, f, n$  are vectors of dimension  $N$ , and  $H$  denotes the PSF block matrix whose elements can be constructed [4,5] from the PSF samples  $\{h(1), h(2), \dots, h(N)\}$ . It should be noted that Equations (3) and (4) represent space-domain models and are equivalent to the frequency-domain model

$$G(\omega) = H(\omega)F(\omega) + N(\omega) \quad (5)$$

where  $\omega$  is the discrete frequency variable and  $G(\omega)$ ,  $F(\omega)$ ,  $H(\omega)$ , and  $N(\omega)$  are DFT's of the  $N$ -point sequences  $g(i)$ ,  $f(i)$ ,  $h(i)$ , and  $n(i)$  respectively.

## 2.2 Image Restoration and Superresolution Problems

For application in missile seeker environments, Equation (3) describes the process of image formation when an unknown object with radiance distribution  $\{f(i)\}$  is imaged through a sensor with a shift-invariant PSF  $\{h(i)\}$ . As noted earlier, practical seeker antenna patterns have a low-pass spectral characteristic and consequently the image obtained is a low-pass filtered version of the object (or scene) being imaged. The problem of interest is then to recover the object, i.e.  $\{f(i)\}$ , by solving Equation (3) (or Equation (5)). However, since the noise sequence  $\{n(i)\}$  will not be known exactly, one will not be able to solve Equation (3) for  $\{f(i)\}$  exactly even when  $\{h(i)\}$ , the PSF of the seeker antenna, is exactly known. One can only hope to obtain an estimate  $\{\hat{f}(i)\}$  which is in some sense close to the original  $\{f(i)\}$ , based on some reasonable assumptions on the noise process  $\{n(i)\}$ . If a distance measure  $J(g, f)$  between  $g$  and  $f$  is used as a norm to measure the closeness of the estimate, the problem of interest can be specified concisely as obtaining the estimate  $\hat{f}^T = [\hat{f}(1), \hat{f}(2), \dots, \hat{f}(N)]$  such that

$$\hat{f} = \arg \min_f J(g, f) = \arg \min_f J\left(g(i) - \sum_j h(i-j)f(j)\right) \quad (6)$$

An examination of the frequency spectra of the object and the image is useful to see clearly the effect of the seeker antenna. Let us assume that the object is space-limited with spatial extent  $\varepsilon$  and, without any loss of generality, assume that  $f_i$  is nonzero only on the interval  $\left[-\frac{\varepsilon}{2}, \frac{\varepsilon}{2}\right]$ . This implies that the spectrum  $F(\omega)$  has infinite extent, i.e. the object has infinite bandwidth, and in the discrete frequency domain, the spectral components in  $F(\omega)$  extend all the way to  $\frac{\omega_s}{2}$ , the folding frequency, as shown in Fig. 1a.

The image spectrum  $G(\omega)$  is a low-pass filtered version of  $F(\omega)$  with the cut-off frequency  $\omega_c$  determined by the diffraction limit of the sensor. Assuming an ideal low-pass filter characteristic, the shape of  $G(\omega)$  will be as shown in Fig. 1b with the spectral components removed in the interval  $\omega_c \leq \omega \leq \omega_s/2$ . The



degradations in the image are hence caused by three factors: (1) spectral mixing within the passband  $0 \leq \omega \leq \omega_c$  due to the convolution with the PSF of the seeker antenna; (2) spectral attenuation caused by removal of spectral

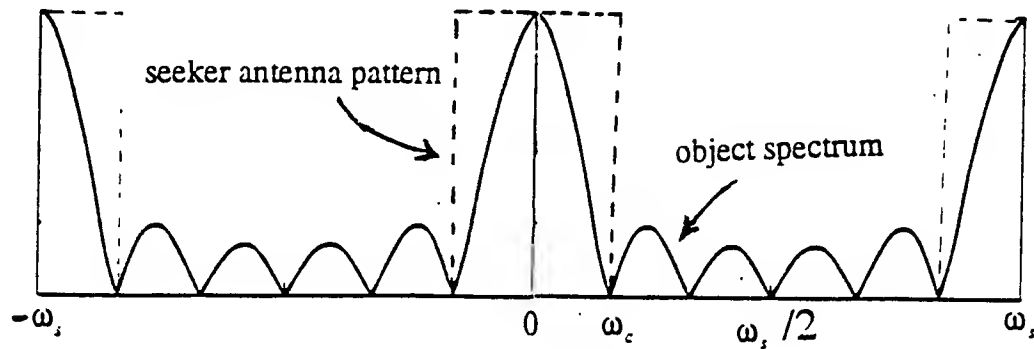


Fig. 1a. Low pass filtering by seeker antenna

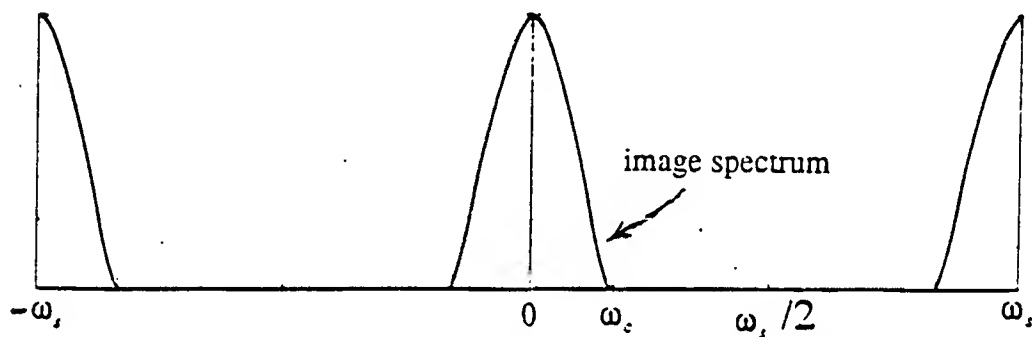


Fig. 1b. Image spectrum resulting from low pass filtering of object spectrum.

components outside the passband; and (3) corruption of the passband due to the additive noise process  $\{n(i)\}$ . Perfect image restoration requires compensation for all three factors cited above.

Traditional image restoration methods attempt mainly at passband reconstruction, i.e. to eliminate the degradations caused by the first and the third factors above. This is achieved by various deconvolution and noise filtering approaches [5,6]. The goal of superresolution is to correct for all three of the above factors, and hence in addition to restoration of spectral components in the passband, extrapolation of the spectrum beyond  $\omega_c$  is to be achieved. Evidently, the ideal of restoring all lost spectral components may be too ambitious and hence realistically one may have to be content with some spectral extrapolation which facilitates recovering the spectrum in the interval  $\omega_c \leq \omega \leq \omega_s$ , where  $\omega_s < \omega_s/2$  is an extended frequency limit. It is of interest to note that even if this limited goal is attained, then the effective cut-off frequency is moved from  $\omega_c$  to  $\omega_s$  and hence the processed image appears as the image acquired from a higher resolution (more expensive, larger aperture) sensor with this larger cut-off frequency. It should also be noted that since generation of new frequency components not present in the original image is attempted, some form of nonlinear processing becomes essential, since linear signal processing methods can not produce frequencies not present in the input signal.

To illustrate the complexity in solving problems of this type, consider the simplest case when no spectral extrapolation is needed, the PSF  $\{h(i)\}$  of the sensor is assumed to be known and the noise  $n_i$  is ignored. This is the classical deconvolution problem [5,6] of solving the vector equation

$$g = Hf \quad (7)$$

for  $f$  given  $g$  and  $H$ , and a solution can be attempted in the form of an "inverse filter" given by

$$\hat{f} = H^{-1}g. \quad (8)$$

Unfortunately, there are several problems with this approach. The system of equations given by equation (7) is often underdetermined which results in  $H^{-1}$  being not defined. Even if  $H^{-1}$  (or a generalized inverse of  $H$ ) can be computed, the estimate  $\hat{f}$  obtained may be worthless due to the presence of noise that was ignored. Observe from the image formation model given by Equation (4), when the presence of noise is accounted for,

$$\hat{f} = H^{-1}g - H^{-1}n.$$

It is now clear that  $H$  being a low-pass filter,  $H^{-1}$  corresponds to a high-pass filter and hence the noise is greatly amplified in the solution estimate. A difficulty of a related nature which also can make the solution given by Equation (8) of limited value is that an exact knowledge of  $H$  is needed for computing the solution and even a small uncertainty in the parameters describing the sensor PSF can result in a very large discrepancy in the solution. In other words, the solution given by Equation (8) is not "robust" enough to tolerate these nonideal conditions that may exist in practice making the estimate obtained useless. Finally, the inverse filter solution is a linear operation and provides no extrapolation of spectrum thus lacking any capability for superresolving. It will be seen later that the drawbacks of this solution procedure stem from the fact that no use of any *a priori* knowledge about the object being restored is made.

The idea of recreating the spectral components that are removed by the imaging process and hence are not present in the image available for processing may pose some conceptual difficulties, which may lead one to suspect whether superresolution is indeed possible. Fortunately there exist sound mathematical arguments confirming the possibility of spectral extrapolation. The primary justification comes from the Analytic Continuation Theorem and the property that when an object has finite spatial extent its frequency spectrum is analytic [7]. Due to the property that a finite segment of any analytic function in principle determines the whole function uniquely, it can be readily proved that knowledge of the passband spectrum of the object allows a unique continuation of the spectrum beyond the diffraction limit imposed by the imaging system. It must be emphasized that the limited spatial extent of the object is critical in providing this capability for extrapolation in the frequency domain.

### 2.3 Use of *a priori* Knowledge in Solution Process

As noted in the last section, due to the ill-posed nature of the inverse filtering problem underlying image restoration and superresolution objectives, it is necessary to have some *a priori* information about the ideal

solution, i.e. the object  $f$  being restored from its image  $g$ . In algorithm development, this information is used in defining appropriate constraints on the solution and/or in defining a criterion for the "goodness" of the solution. How to utilize this information is at the heart of a well-tailored superresolution algorithm.

The specific *a priori* knowledge that can be used evidently depends on the specific application. For applications in astronomy, it could come in the form of some known facts about the spectral differences of the objects one is looking for (for instance, a double star as opposed to a star cluster). In medical imaging and in military applications, it could come from the geometrical features of the object (target shape, for instance). For radar and MMW imagery, one could use the fundamental knowledge that the reflectivity of any point on the ground can not be negative. In addition to the nonnegativity constraint, a space constraint resulting from the known space-domain limits on the object of interest could be used. Other typically available constraints include level constraints (which impose upper and lower bounds on the intensity estimates  $\hat{f}_j$ ), smoothness constraints (which force neighboring pixels in the restored image to have similar intensity values) and edge-preserving constraints. More complicated constraints are possible, but in general they result in tuning the algorithms to specific classes of targets.

Varying by the extent to which *a priori* knowledge can be incorporated in algorithm development, there have been introduced into the literature a large number of image restoration approaches and algorithms too vast to describe or reference here. One may refer to some recent survey papers [8,9] for a review of the extensive activity on this topic. In this section, we shall only briefly cite a few of the approaches that have received some interest in the context of superresolution capabilities, i.e. those that provide possible spectral extrapolation. It should be noted clearly that not all image restoration methods provide the capability for superresolving. In fact, a majority of existing schemes may perform decent passband restoration, but provide no bandwidth extension at all.

The various approaches in general attempt to code the *a priori* knowledge to be used by specifying an object model or a set of constraint functions, and further employ an appropriate optimization criterion to guide in the search for the best estimate of the object. A convenient way of classifying the resulting algorithms is into **iterative** and **noniterative** (or direct) schemes. Noniterative approaches generally attempt to implement an inverse filtering operation (without actually performing the computation of the inverse of the PSF matrix  $H$ , however) and have poor noise characteristics. All required computations and any possible use of constraint functions are applied in one step. In contrast, iterative methods apply the constraints in a distributed fashion as the solution progresses (as shown in Fig. 2) and hence the computations at each iteration will be generally less intensive than the single-step computation of noniterative approaches. Some additional advantages of iterative techniques are that, (1) they are more robust to errors in the modeling of the image formation process (uncertainties in the elements of the PSF matrix  $H$ , for instance), (2) the solution process can be better monitored as it progresses, (3) constraints can be utilized to better control the effects of noise (and possibly clutter), and (4) can be tailored to offset sensor nonlinearities. The disadvantages of these methods generally are, (1) increased computation time, and (2) need for proving convergence of the iterative scheme (in fact, for some algorithms this could be impossible). In

the development of an efficient processing algorithm for a specific application one needs to evaluate these tradeoffs and tailor the steps of the algorithm to exploit the inherent features available in that application.

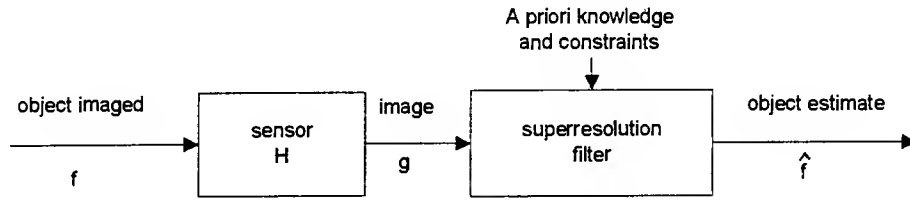


Fig. 2a. Schematic of Noniterative (Direct) Superresolution

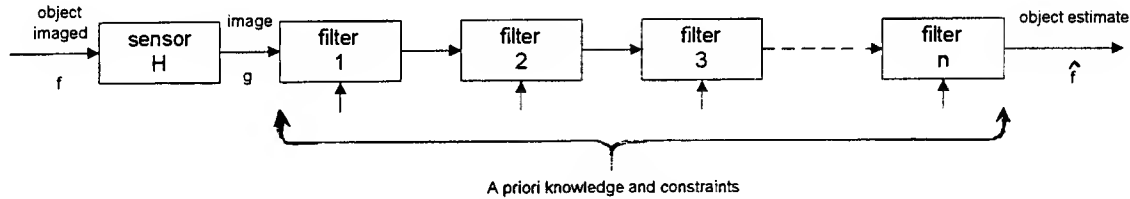


Fig. 2b. Schematic of iterative Superresolution

### 3. SPECIFIC ALGORITHMS FOR IMAGE RESTORATION AND SUPERRESOLUTION

Among the several approaches that utilize iterative or noniterative processing techniques, a few specific algorithms have been receiving a greater share of attention in regard to their claims for restoration and superresolution performance when applied to multispectral seeker data. During this project, we focused our attention on two candidate algorithms, one of which is a representative of the noniterative class of algorithms and the other being a representative of the iterative class. A brief outline of these algorithms which is followed by some qualitative and quantitative performance evaluations that will highlight the strengths and weaknesses of the two processing approaches will be given in this section.

#### 3.1 A Noniterative Algorithm Using Matrix Computations

The development of noniterative algorithms that employ simple matrix operations has been a popular line of investigation in recent times. One of the more well known algorithms of this type has been given by Gleed and Lettington [10] using a regularized pseudo-inverse computation approach. Starting with the space-domain image formation model given by Equation (4), Gleed and Lettington note that evaluating the solution as

$$\hat{f} = H^{-1}g - H^{-1}n \quad (9)$$

provides a poor quality estimate due to the noise amplification caused by  $H^{-1}$  (in turn due to some eigenvalues of  $H$  becoming too small). To overcome this difficulty, they propose to modify the estimate by first diagonalizing the  $H$  matrix through the transformation

$$M^T H M = \Lambda$$

where  $M$  is the modal matrix of  $H$  and  $\Lambda$  is a diagonal matrix with the eigenvalues of  $H$  along the diagonal [11]. The object estimate  $\hat{f}$  is then obtained as

$$\hat{f} = H_{\text{mod}}^{-1}g - H_{\text{mod}}^{-1}n \quad (10)$$

where  $H_{\text{mod}}^{-1}$  is computed as

$$H_{\text{mod}}^{-1} = M[\Lambda + \mu_1 \Lambda^{-1}]M^T. \quad (11)$$

$\mu_1 \geq 0$  is a scalar parameter to be selected appropriately based on the noise present  $n$ .

The solution given by (11) changes the PSF of the imaging system from  $H$  to  $H_{\text{mod}}$ , however, and it is necessary to account for this change. Gleed and Lettington [10] propose a “regularization” operation by constructing the matrix  $R$  as

$$R = H_{\text{mod}}^{-1}H + \mu_2(H_{\text{mod}}^{-1}H)^{-1} \quad (12)$$

and obtaining the final estimate as

$$\hat{f} = R^{-1}H_{\text{mod}}^{-1}g. \quad (13)$$

In Equation (12),  $\mu_2$  is another user selected parameter satisfying the condition  $\mu_2 \leq \mu_1$ . Gleed and Lettington [10] report getting satisfactory resolution improvements in processing various images including PMMW imagery. The exact extent of spectral extrapolation obtained by this method is however not clear. Furthermore, the selection of scalars  $\mu_1$  and  $\mu_2$  is rather ad hoc.

### 3.2 A Maximum Likelihood (ML) Algorithm for Iterative Superresolution

An important class of algorithms that are receiving particular attention in recent times for their superresolution capabilities are those that can be developed starting with a statistical modeling of the imaging process. The basic idea underlying these methods is to account for the statistical behavior of emitted radiation at the level of individual photon events by constructing appropriate object radiance distribution models (using knowledge of fluctuation statistics).

For a brief description of these methods, let  $f(x)$  denote the object's intensity function,  $x \in X$ , where  $X$  defines the region over which intensity is defined, and let  $g(y)$  denote the intensity detected in the image,  $y \in Y$ , where  $Y$  defines the region over which intensity is detected. If  $\{h(y, x), y \in Y \text{ and } x \in X\}$  denotes the point spread function (PSF) of the imaging sensor, then accounting for the presence of noise in the imaging process, one can model the imaging process by

$$g(y) = \sum_{x \in X} h(y, x)f(x) + \text{noise} \quad (14)$$

(where an additive noise is assumed as before for the sake of simplicity). The classical restoration problem is to find the object intensity estimate  $\{\hat{f}(x)\}$  given the data  $\{g(y)\}$ .

There exists considerable literature on developing explicit algorithms, mainly of an iterative nature, for handling the image restoration problem within a statistical framework afforded by such a formulation. A particularly attractive approach is to obtain a maximum likelihood (ML) estimate  $\{\hat{f}(x)\}$  i.e. the object intensity estimate that most likely have created the measured data  $\{g(y)\}$  with the PSF process  $\{h(y,x)\}$ , which in turn is developed by maximizing an appropriately modeled likelihood function (or the logarithm of this function, for simplicity). Modeling the likelihood function is basically obtaining a goodness-of-fit (GOF) quantity for the measured data, since the likelihood function is a statistical distribution function  $p(g/f)$  obtained as a fit to the relation between the data  $\{g(y)\}$  and the object  $\{f(x)\}$ . The success of image restoration in a given application depends on how good the assumed conditional probability function fits the input/output characteristics of the imaging system. While a commonly used model is a Chi-squared function, very active and intense research continues to this day on the development of improved statistics that lead to better ML estimates.

In the formulation of iterative algorithms that afford simple implementation, an important contribution has been the work of Shepp and Vardi [12] who used the Expectation Maximization (EM) algorithm (originally suggested by Dempster et. al. [13]) to solve positron emission tomography imaging problems in which Poissonian statistics are dominant. The major advantage of using the EM algorithm is that it involves the solution of linear equations, whereas the original ML problem is in general a nonlinear optimization problem. Following this approach, Shepp and Vardi [12] developed an iterative algorithm for which convergence can be proven analytically. This algorithm also reduces to the familiar Richardson-Lucy iteration which has attained considerable popularity in the fields of astronomy and medical imaging. For a discretized formulation of the imaging equation (2), with  $g(j)$  and  $f(j)$ ,  $j = 1, 2, \dots, N$ , denoting the  $N$  pixels of the image and the object respectively, and  $h(j)$  denoting the PSF of sensor, the updating of the object estimates takes the form

$$\hat{f}_{k+1}(j) = \hat{f}_k(j) \left[ \frac{g(j)}{\hat{f}_k(j) \otimes h(j)} \otimes h(j) \right], \quad j = 1, 2, 3, \dots, N, \quad (15)$$

where  $k$  denotes the iteration count and  $\otimes$  denotes discrete convolution. The initial estimate  $\hat{f}_0(j)$  is taken as the image  $g(j)$  to commence the iteration.

It should be noted that the ML estimate  $\hat{f}(j)$  attempts to construct an estimate for  $f(j)$ , the number of photons emitted by the  $j$ -th sample of the object (which is considered a random variable) from a knowledge of  $g(j)$ , the  $j$ -th pixel value in the input image, and  $h(j)$ , the  $j$ -th element of the sensor PSF. The optimization

framework in which the algorithm is developed ensures that the likelihood function  $p(g/f)$  monotonically increases over successive iterations of the algorithm and hence the processed image improves in quality as the algorithm processed.

### 3.3 Comparison of Algorithms for Multispectral Seeker Implementations

Two distinct algorithms for restoration and superresolution of image data that employ iterative and noniterative computations were outlined in the last two sections. Due to the differences underlying the approaches followed in obtaining these algorithms, they possess particular advantages and disadvantages. For an intelligent selection of the right approach in a specific seeker application, one needs to weigh these advantages and disadvantages in the light of some basic requirements that need to be met in these implementations. These are listed in the following:

1. Flexibility for application to images from different sensing modalities;
2. Performance robustness to tolerate modeling uncertainties, parameter inaccuracies and nonlinearities;
3. Computational requirements that can be met in typical real-time applications;
4. Ensure desired level of resolution enhancement in the presence of significant noise levels;
5. Ensure satisfactory performance in realistic clutter scenarios (with signal-to clutter ratios (SCR) in the range 5-10dB)

The limits on complexity and computational requirements are evident, given the real time operation requirement for a missile seeker. On the surface it may appear that noniterative approaches provide an obvious advantage over iterative approaches on this count since the entire computation needs to be performed only once. However, there are several factors that need to be considered in this evaluation. First of all, the computation required for implementing the noniterative estimation given by (13) is considerably more complex than the implementation of the iteration given by (15). It must be emphasized that the discrete model for the imaging process, viz, Equation (4), which is used as the basis for the noniterative estimation results in a matrix  $H$  of typically large dimension. Consequently, the matrix inversion operations required in (18) can pose some difficulties. Furthermore,  $H$  will have a number of zero elements, which further add complexities to the computation of inverses. In our implementations we have found that execution of iterations of the form given by (15) over several cycles is in many cases computationally preferable to the noniterative algorithm (18) that requires matrix inverse calculations.

Perhaps a greater advantage of iterative algorithms of the type (15) comes from the performance robustness to noise, clutter and parameter uncertainties. It is evident that an accurate knowledge of the sensor PSF matrix  $H$  is necessary for computation of the estimates in (13), whereas a significant tolerance to inaccuracies in the sensor PSF is provided by the ML iteration algorithm (15). In fact, this algorithm can be modified, as will be described in the next section, for a blind implementation when a complete knowledge of the sensor PSF is not initially available and one would like to obtain improved estimates of the PSF parameters as the iterations progress.

The requirements arising from the presence of noise and clutter are also evident from the practical environments in which target detection and classification are to be performed. As noted earlier, there are two main

sources of noise in these applications, viz, the receiver noise, whose statistics depend on the type of imaging sensor employed and are usually signal dependent, and the quantization noise, which can be realistically modeled by a zero-mean white Gaussian random field that is independent of the image signal. It is well known that deconvolution methods (particularly those that attempt to implement directly an inverse filter) are highly sensitive to noise and require rather high SNR levels for satisfactorily processed images. One may note that typical radiometric images (PMMW images, for instance) have SNR levels of about 20dB (or less), thus highlighting the importance of this requirement.

Finally, it is beneficial to have the restoration algorithm be capable of processing signals collected from different types of sensors. This is due to the fact that present day missile seekers are required to handle data collected from different sensing modalities (typically operating in different frequency ranges) and to perform tactical decision-making based on the fused data. In these environments, any signal processing on the measured data aimed at contrast enhancement [14] and/or resolution enhancement (superresolution) often comes as a first step operation which facilitates the further processing steps, such as feature extraction, feature integration for fusion, etc. [15], to be implemented more efficiently. Use of iterative algorithms that have certain inherent robustness to parameter inaccuracies can provide major benefits since characterization studies leading to accurate determination of PSF may not be readily available for every detector in the sensor suite.

The qualitative comparisons given in this section favor the selection of iterative algorithms of the type given by (15). While convergence of iterative schemes is in general an issue of concern, analytical support ensuring the convergence of the ML iterations in (15) is available. In practice, one can terminate the iterations once a desired resolution level in the processed image is attained. Development of additional constructs that speed up the convergence of the ML iterative algorithm given by (15) is a topic that is receiving considerable attention.

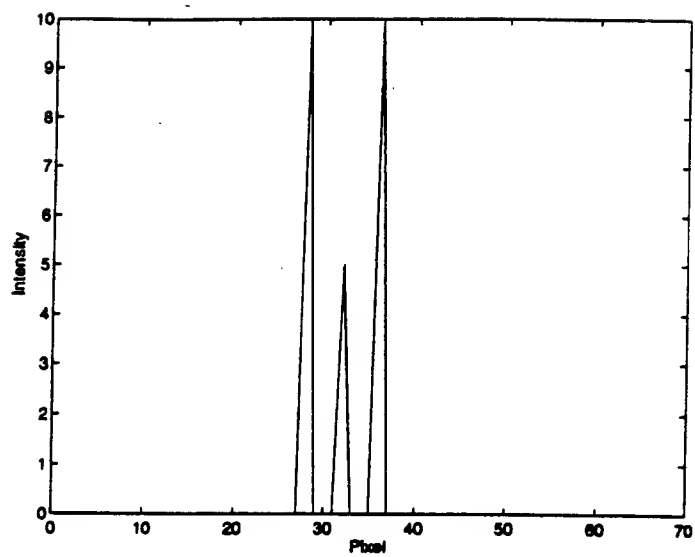
### **3.4 Quantitative Performance Evaluation of ML Superresolution Algorithm**

Several experiments directed to evaluating the restoration and superresolution performance of the ML iterative algorithm given by (15) were conducted as part of this project. Results from four experiments are briefly summarized in this section to illustrate the efficiency of this algorithm in the processing of various types of input signals. Experiments 1-3 were principally aimed at measuring the spectrum extrapolation performance and hence were conducted in a "controlled environment" by starting with a known object (1-D or 2-D) and blurring it with a known PSF to generate the input signal for the processing algorithm. On the other hand, experiment 4 dealt with evaluating the performance in processing "real data", which in this case is a passive MMW image collected by Wright Laboratory Armament Directorate personnel at Eglin AFB using a state-of-the-art passive MMW data acquisition platform comprising of a 95 GHz 1 foot diameter aperture radiometer.

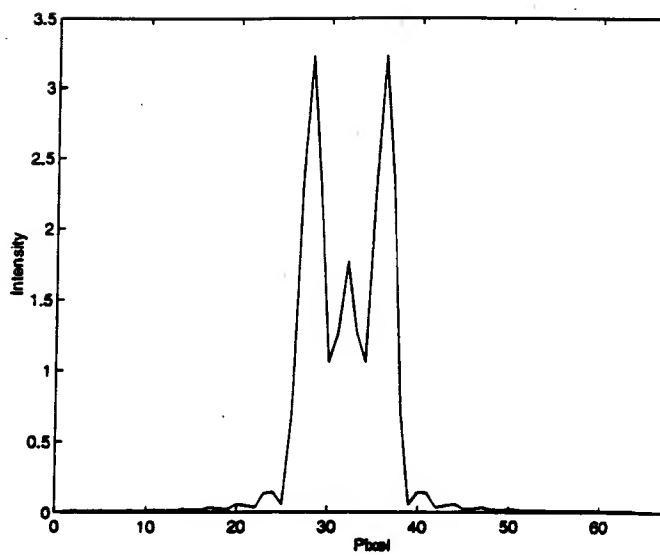
#### **Experiment 1:**

Fig. 3a shows the original object comprising of 3 impulses (point sources) located 2 pixels apart. Fig. 3b shows the image formed when blurred with a sensor with a cutoff frequency 21. Figs. 3c and 3d give the reconstructed images after 90 and 1000 iterations of the ML algorithm. While the restoration of the original object

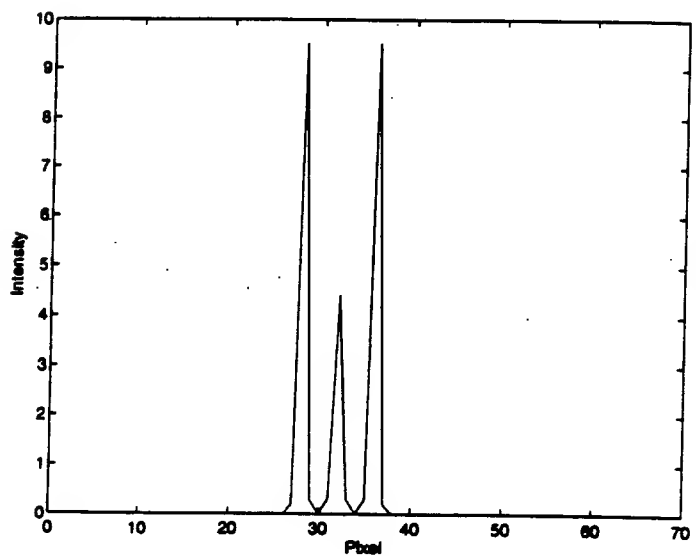




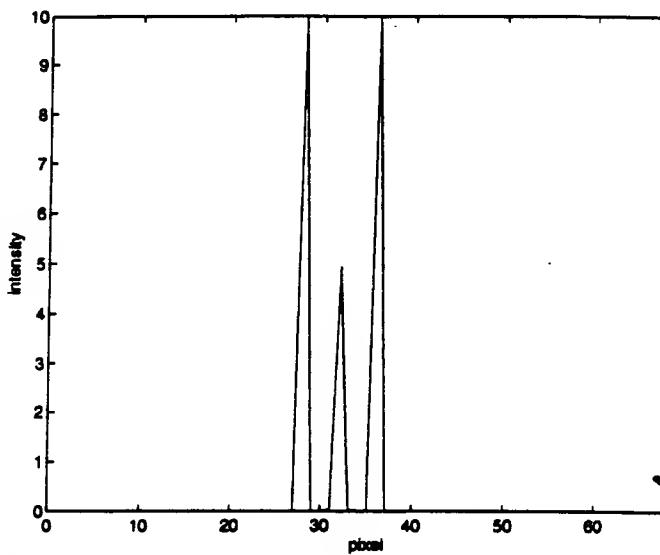
3a



3b



3c



3d

Fig. 3. Signal restoration performance in Experiment 1

is clearly seen in these figures, the spectrum extrapolation ( and hence superresolution ) can be more clearly seen by examining the frequency spectra of the signals before and after processing. Fig. 4a shows the spectrum of the original object, Fig. 4b shows the spectrum of the image formed ( note the cutoff frequency 21 of the sensor ) and Fig. 4c shows the spectrum of the reconstructed image after 90 iterations of the ML algorithm. The extrapolation of frequency components beyond the cutoff at 21 clearly demonstrates that this algorithm is superresolving.

#### **Experiment 2:**

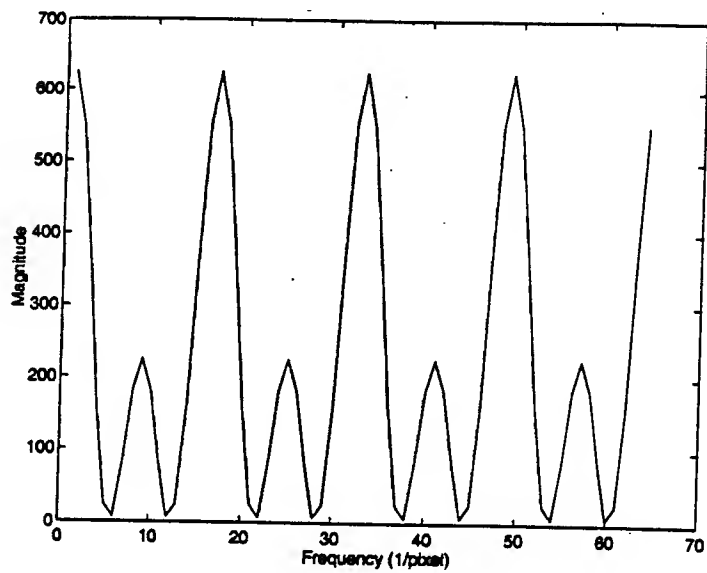
Fig. 5a shows a more complex one-dimensional object characterized by several edges and hence offers a greater challenge to the processing algorithm. The image formed by blurring with a sensor with a cutoff frequency 21 is shown in Fig. 5b. Fig. 5c shows the reconstructed image after 90 iterations of the ML algorithm, where the edge reconstruction is clearly seen. The spectrum extrapolation performance can also be seen by comparing Fig. 6a, 6b and 6c, which show the spectra of the original object, input image and the ML processed image (after 90 iterations). It should be noted that only the portion of the spectrum in the high frequency range is plotted to an expanded scale in these figures since the signal has low frequency components with relatively large magnitudes that prevent the high frequency components from being displayed effectively on the same graph. For testing the restoration performance in the presence of noise, another experiment was conducted by blurring the same object as before and then corrupting with an additive Gaussian noise to result in an image with a Signal-to-Noise Ratio (SNR) of 30.0248 dB shown in Fig. 7a. The reconstructed image after 200 iterations of ML algorithm is shown in Fig. 7b which confirms the noise filtering properties of the algorithm.

#### **Experiment 3:**

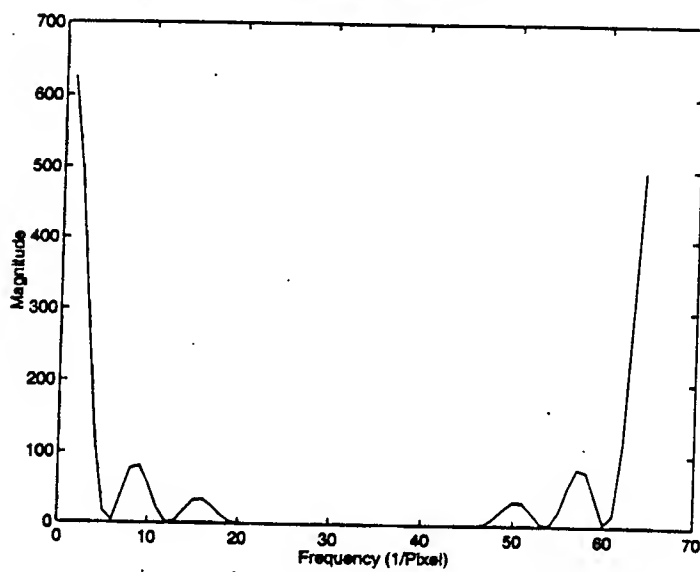
The superresolution performance of these algorithms was also tested by processing a 256 x 256 image from our database. Fig. 8a shows the original image (Lena) used in this experiment. Fig. 8b shows the blurred image obtained by convolution with the PSF of a sensor with cutoff  $\omega_c = 63$  (which is approximately one-half of the folding frequency and is typical of practical imaging operations). The restoration performance of ML algorithm as the number of iterations is gradually increased is shown in the next set of figures (Figs. 8c - 8f where the restored images after 10, 20, 90, and 100 iterations are shown). The resolution enhancement with only 10 iterations of the algorithm is clearly visible. The algorithm was stopped at the end of 100 iterations since the resolution is comparable to that of the original (unblurred) image. Figures 9a - 9d show the power spectra of various signals; the frequency extrapolation performed by the ML algorithm to achieve superresolution is clearly noticeable by comparing the four corners in Figs. 9c and 9d which show the power spectra of the blurred image and the reconstructed image after 100 iterations of ML.

#### **Experiment 4:**

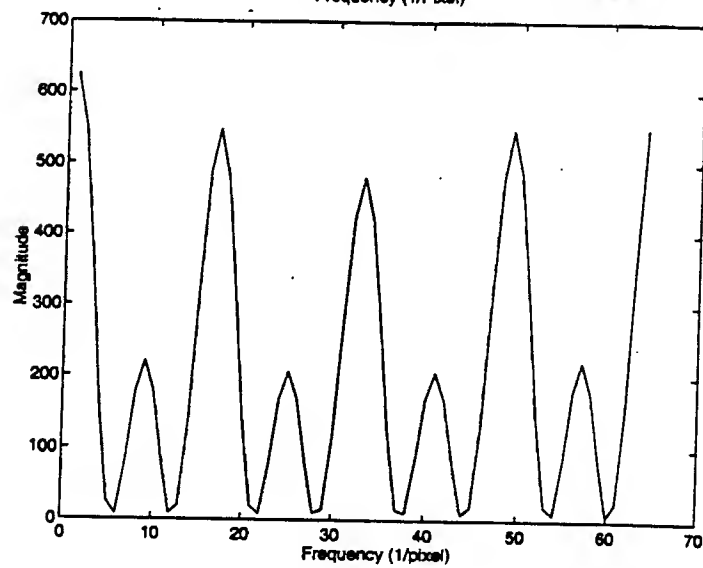
The performance of the ML algorithm was also tested by processing a set of passive MMW images supplied by the Wright Laboratory Armament Directorate personnel. These images were collected under various conditions (time-of-day, atmospheric conditions, etc.), which were not known save for the fact that the images were obtained by a single detector radiometer with 1 foot aperture at 95 GHz. For illustration purposes, the result of processing one of these images ("Jeep 2" image) will be presented here.



4a

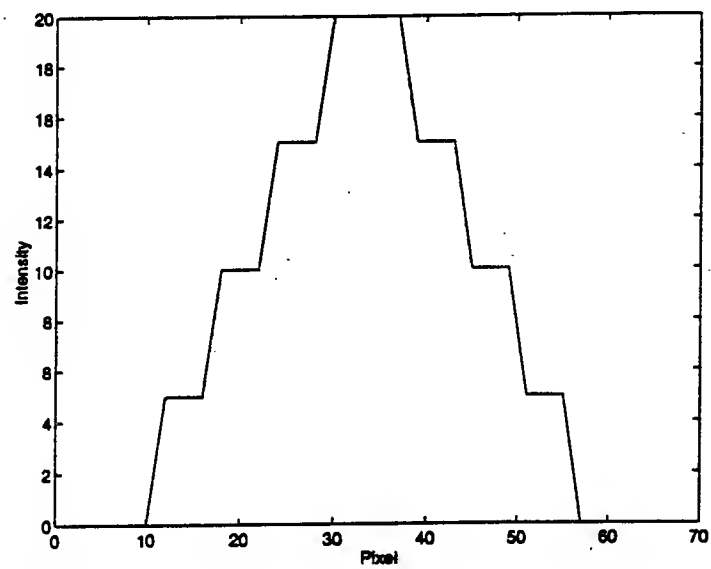


4b

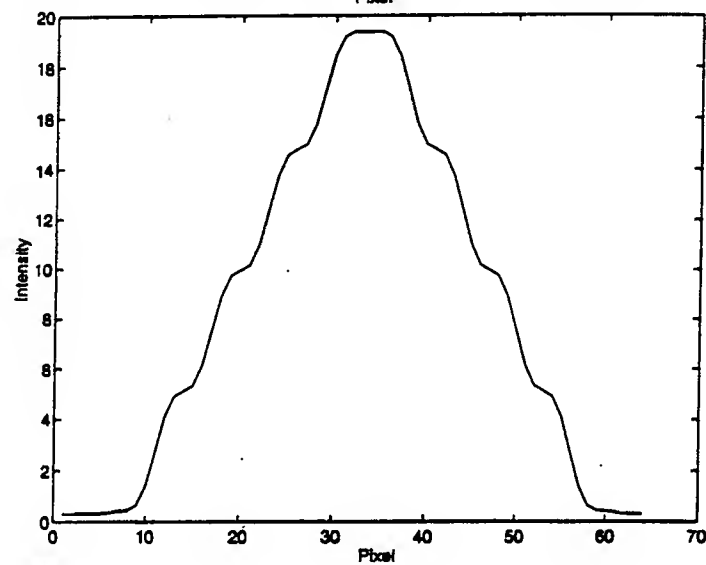


4c

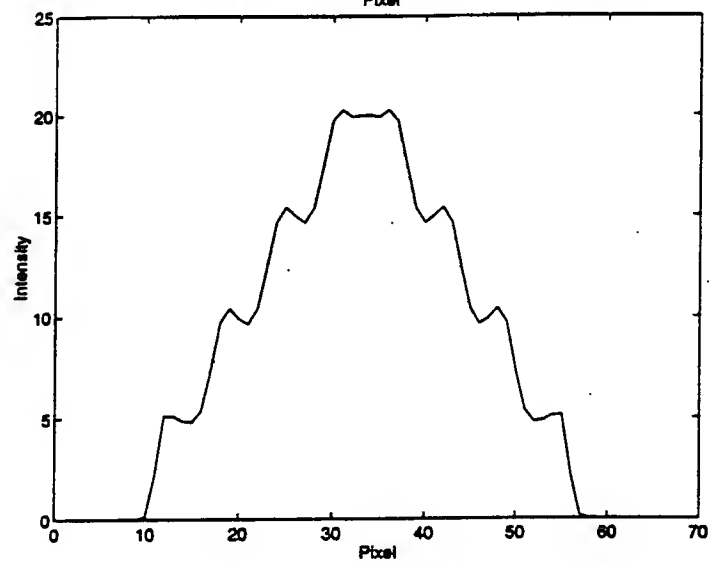
Fig. 4. Spectrum reconstruction performance in Experiment 1



5a



5b



5c

Fig. 5. Signal restoration performance in Experiment 2

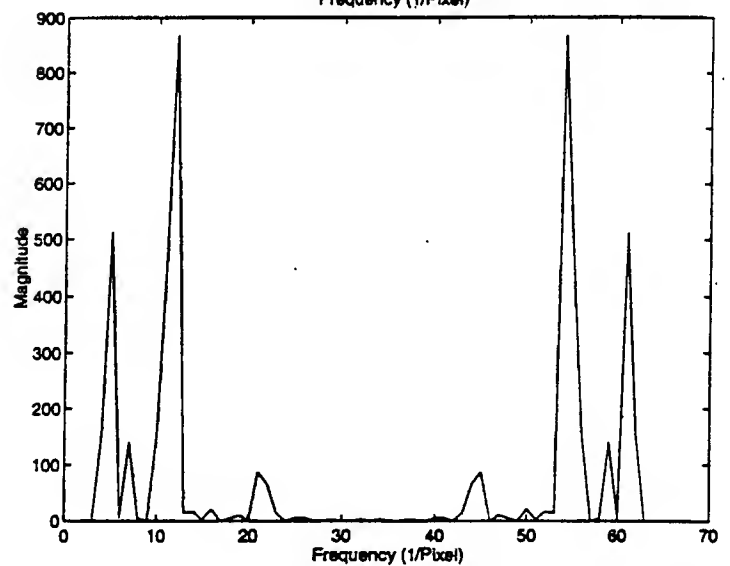
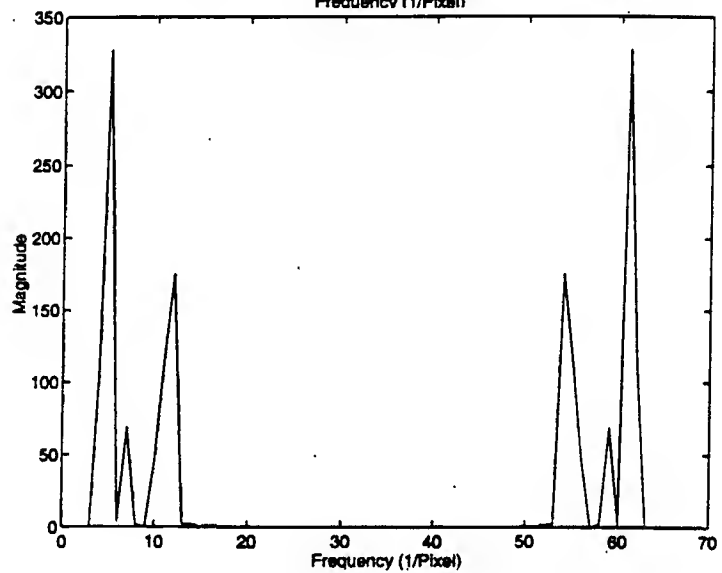
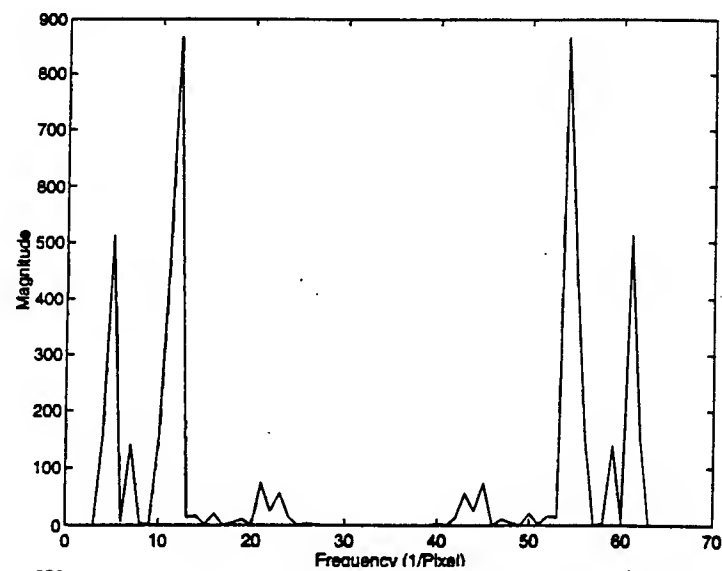
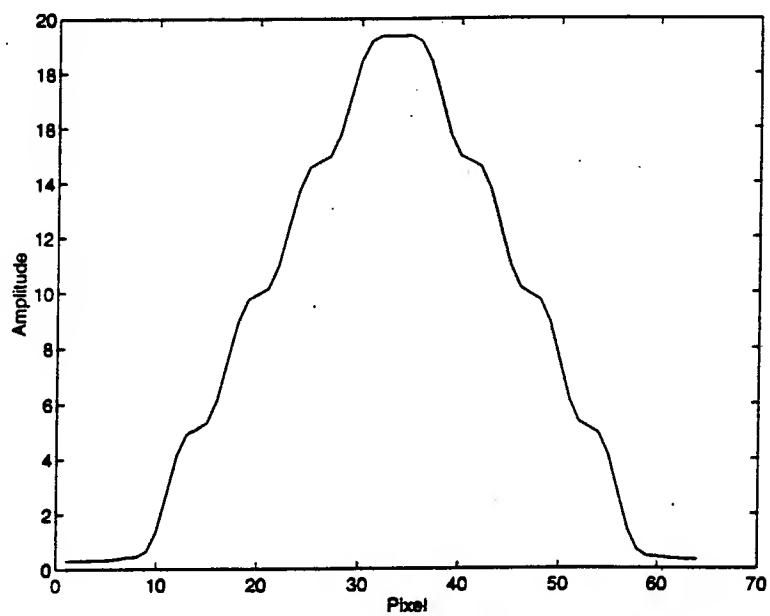
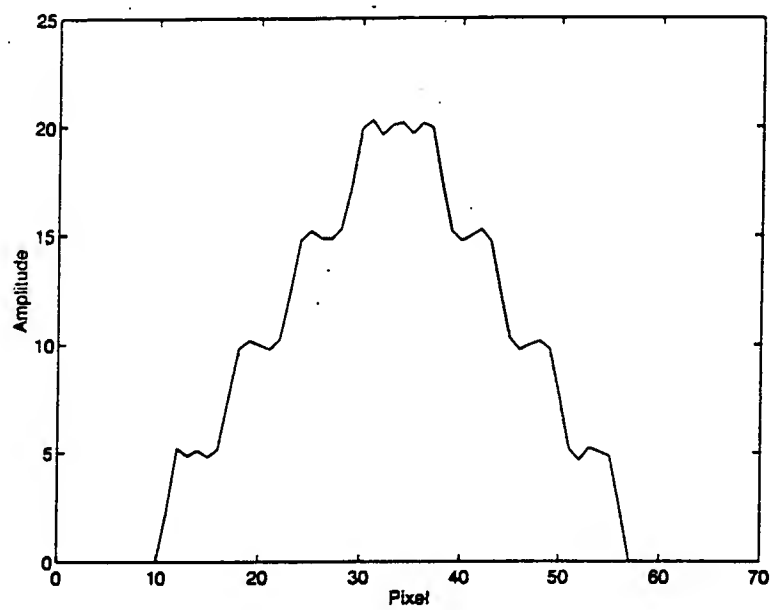


Fig. 6. Spectrum reconstruction performance in Experiment 2

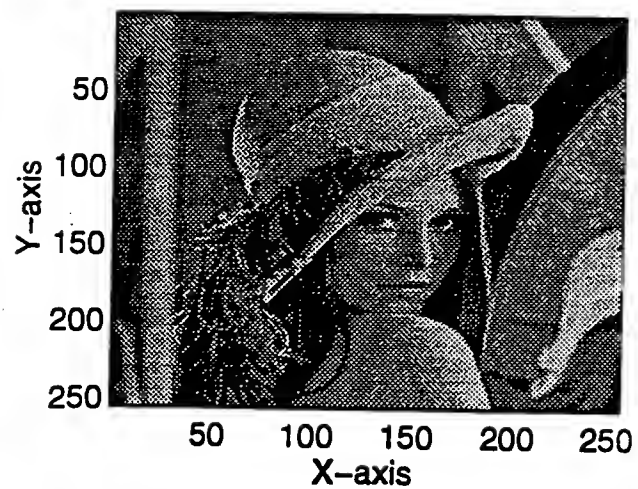


7a

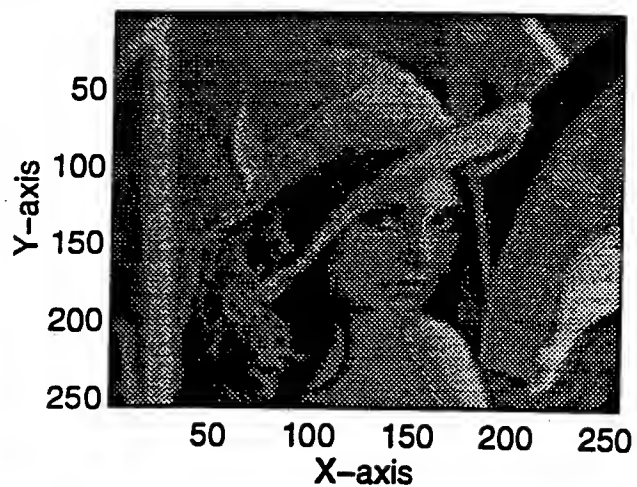


7b

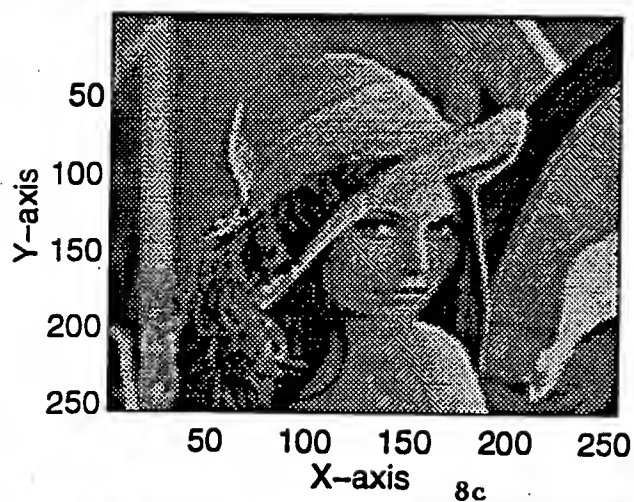
Fig. 7. Signal restoration in the presence of noise in Experiment 2



8a



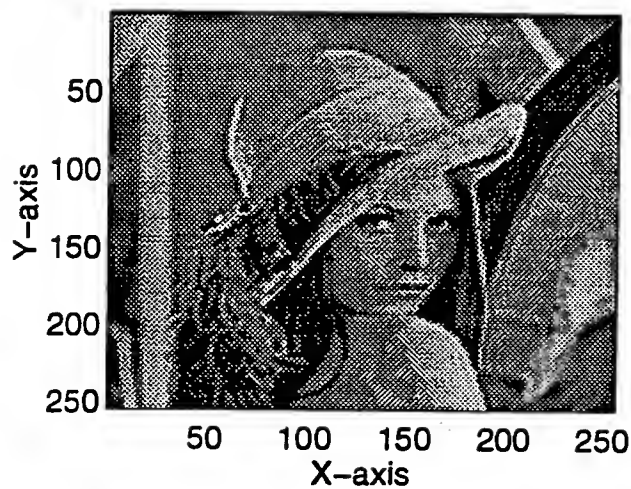
8b



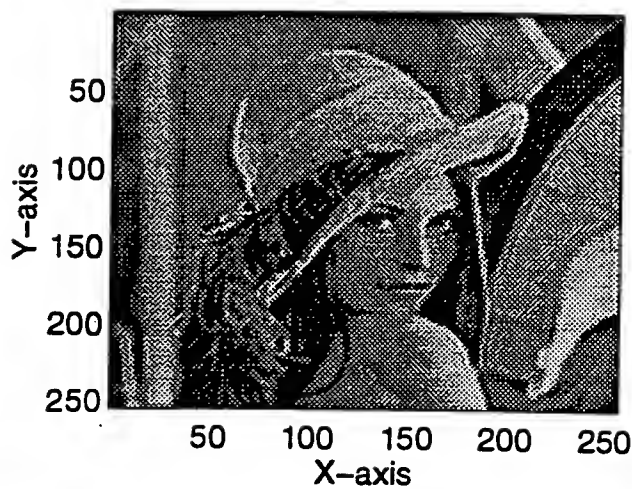
8c



8d



8e



8f

Fig. 8. Image restoration performance in Experiment 3

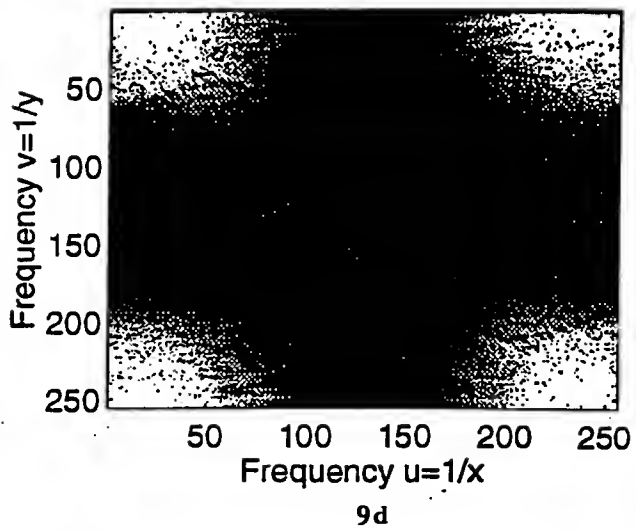
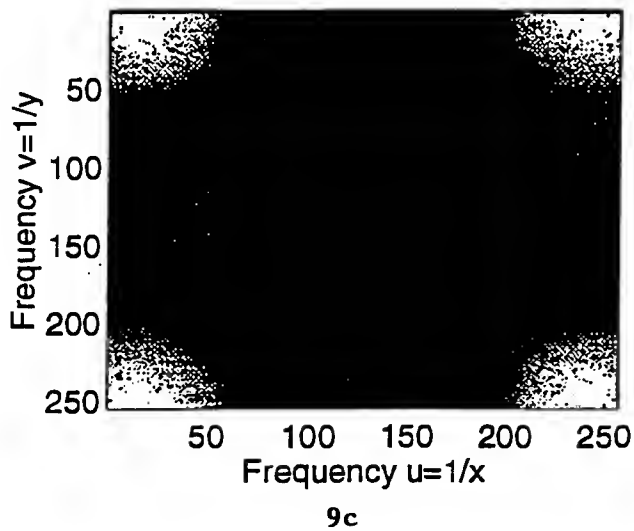
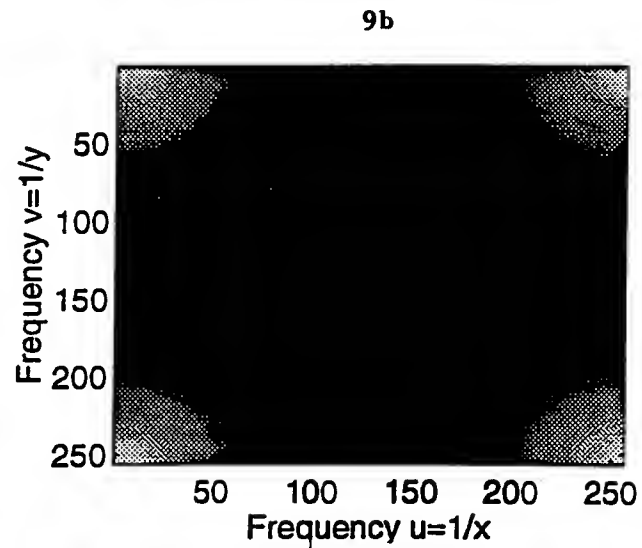
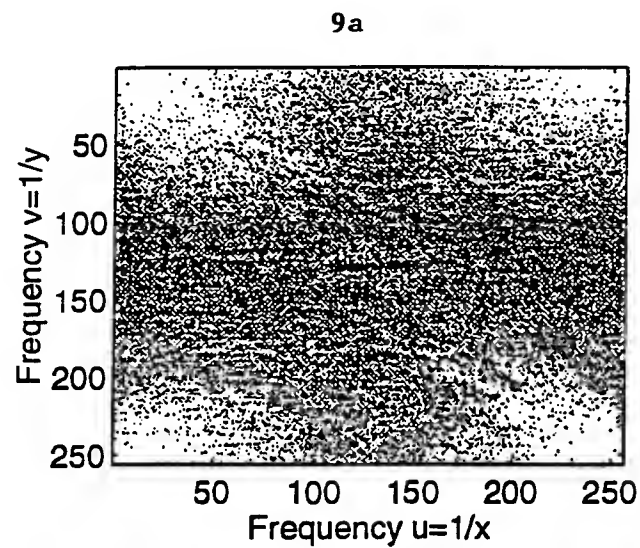


Fig. 9. Spectrum extrapolation performance in Experiment 3



The received image to be processed is shown in Fig. 10a. It is evident that the resolution present is quite poor and the extraction of any useful features from this image is rather doubtful. The image can benefit from further processing aimed at resolution enhancement. Since a complete characterization of the sensor and the imaging conditions were not available to determine the sensor PSF, an approximate analysis was conducted to identify the cut-off frequency in the optical transfer function (OTF). Starting with an intensity profile corresponding to an edge of the object in the image (as shown in Fig. 11a which gives the intensity profile in column 100), an OTF was created as a low-pass filter whose convolution with an input pulse object (shown in Fig. 11b) yields an outcome (shown in Fig. 11c) which is similar to the edge profile. The cutoff in the OTF in this case was estimated to be 17. While this process gives only a rough estimate of the OTF of the imaging system, through iterative adjustments it is possible to obtain a sufficiently accurate characterization of the sensor PSF to commence the ML iterations for superresolution.

Figs. 10b - 10g show the results of processing the input image in Fig. 10a after 10, 20, 30, 40, 60, and 100 cycles of ML iteration. The progressive enhancement of resolution is clearly evident from the strengthening of edges and the improved structural details of the object (particularly near the wheels, the windows and the top of the vehicle).

#### 4. PERFORMANCE IMPROVEMENT BY ITERATIVE BLIND ML RESTORATION

As discussed earlier, if the sensor is fully characterized and if the images contain good target/scene metrics such as exact time of day, distances to objects, weather conditions etc., one may attempt to model the sensor PSF exactly and employ it for restoration of the image [16]. When the PSF parameters are not readily available, one may attempt to build an approximation to the PSF from the image to be processed using techniques such as the one described in the last section (of looking for a column or row in the image corresponding to a sharp edge and matching this profile with the blurred version of a sharp edge passed through an approximately tailored OTF). Since the performance of any deconvolution procedure improves with the availability of accurate PSF parameters, it is useful to consider implementations of iterative restoration and superresolution algorithms where the PSF estimates can be adaptively updated along with the iterative construction of the object estimate. Such implementations can be regarded as special cases of "blind deconvolution" algorithms which attempt to perform image restoration with incomplete knowledge of both the PSF and the object (i.e.  $h$  and  $f$  in the imaging process model (14)). In this section we shall present one such implementation and discuss its performance when applied to superresolution of PMMW images.

##### 4.1 ML Algorithm for Joint Estimation of Object and PSF

For the presentation of the new algorithm, it is useful to briefly review the mathematical basis underlying the iterative ML scheme given by (15). Starting with the imaging equation (14) and assuming that the additive noise can be modeled as an independent and identically distributed (i.i.d.) random variable with a zero mean



10a



10b



10c



10d



10e



10f



10g

Fig. 10. Restoration of passive millimeter-wave image ("Jeep 2")

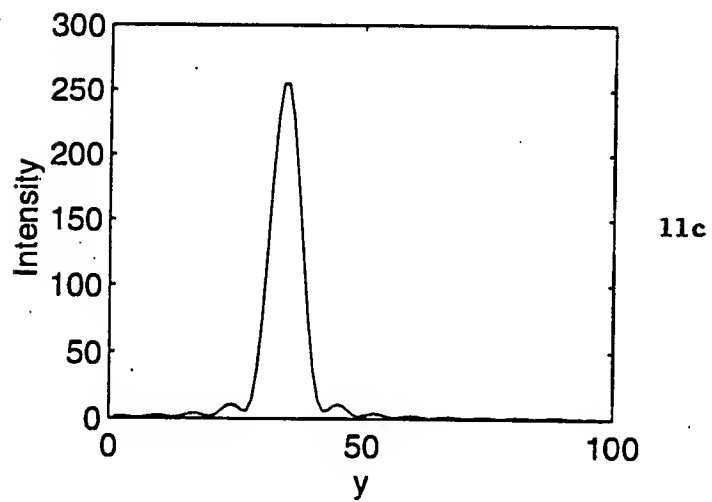
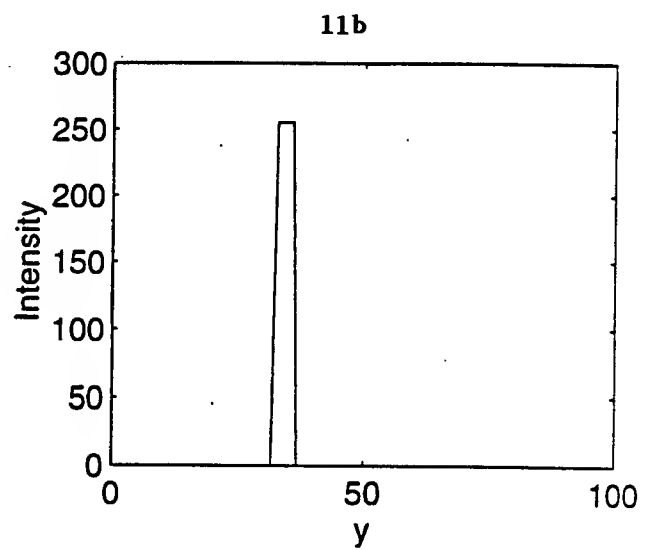
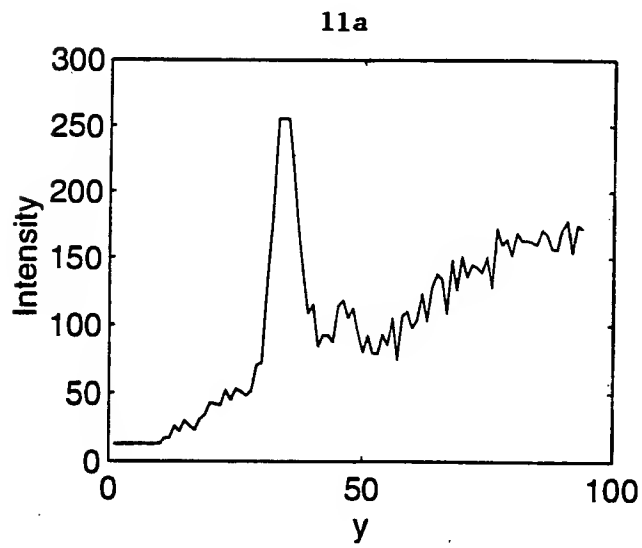


Fig. 11. Estimation of sensor PSF from image

Gaussian probability density having variance  $\sigma_n^2$ , the intensity detected in the image  $g(y)$  can be regarded as a random variable with a normal probability density

$$p[g(y), h, f] = \frac{1}{(2\pi\sigma_n^2)^{\frac{1}{2}}} \exp \left[ -\frac{\{g(y) - \sum_{x \in X} h(y, x)f(x)\}^2}{2\sigma_n^2} \right]. \quad (16)$$

It may be noted that the noise model assumed is particularly appropriate if the dominant noise is thermal noise. The probability density for realizing the entire data set,  $\{g(y), y \in Y\}$  can hence be obtained as

$$P[g, h, f] = \prod_{y \in Y} p[g(y), h, f] \quad (17)$$

which serves as a model for the likelihood function for evaluating the ML estimate  $\{\hat{f}(x)\}$  that most likely have produced the data  $\{g(y)\}$  from  $\{h(y, x)\}$  and  $\{f(x)\}$ . It is obtained by maximizing  $P$  in (17).

For simplicity in computation, one conducts the maximization of a modified log-likelihood function

$$L[g, h, f] = -\sum_{y \in Y} \left[ g(y) - \sum_{x \in X} h(y, x)f(x) \right]^2 \quad (18)$$

which is obtained by taking the natural logarithm of  $P$  in (17) and removing a constant term that does not affect the maximization process. If  $L$  is maximized with respect to  $f$  assuming  $h$  to be known, we have the standard ML restoration, whereas if  $L$  is maximized with respect to both  $f$  and  $h$  by searching over a larger parameter space, we have blind ML restoration.

The iterative scheme given in (15) provides a discretized implementation of the standard ML restoration process. It can be easily converted into a blind restoration scheme by observing that  $\hat{f}(j)$  and  $h(j)$  are interchangeable in the two convolution operations in (15), and hence an updating scheme for  $h(j)$  can be developed for estimation of the PSF along with the object estimation, i.e. obtaining  $\hat{f}(j)$ . For a brief description, each cycle of this "Blind ML restoration algorithm" consists of executing the two steps:

**Step 1:** Implement "object estimation" algorithm through  $m$  iterations with initial guess for  $h_k(j)$  and

$$\hat{f}_o(j) = g(j):$$

$$\hat{f}_{k+1}(j) = \hat{f}_k(j) \left[ \left\{ \frac{g(j)}{\hat{f}_k(j) \otimes h_k(j)} \right\} \otimes h_k(j) \right] \quad (19)$$

**Step 2:** Implement “PSF updating” algorithm through  $n$  iterations:

$$h_{k+1}(j) = h_k(j) \left[ \frac{g(j)}{h_k(j) \otimes \hat{f}_{k+1}(j)} \right] \otimes \hat{f}_{k+1}(j). \quad (20)$$

The algorithm hence continuously reshapes the PSF to result in improved object estimation after each cycle of implementation. A flow-chart depicting the various steps is shown in Fig. 12. The algorithm can be run iteratively over several cycles until a specified maximum iteration count is reached or a processed image with a satisfactory resolution level is attained. The quality of estimation depends on the number of object estimation iterations  $m$  and PSF updating iterations  $n$  implemented within each cycle. Some results describing the performance of this algorithm will be outlined in the next section.

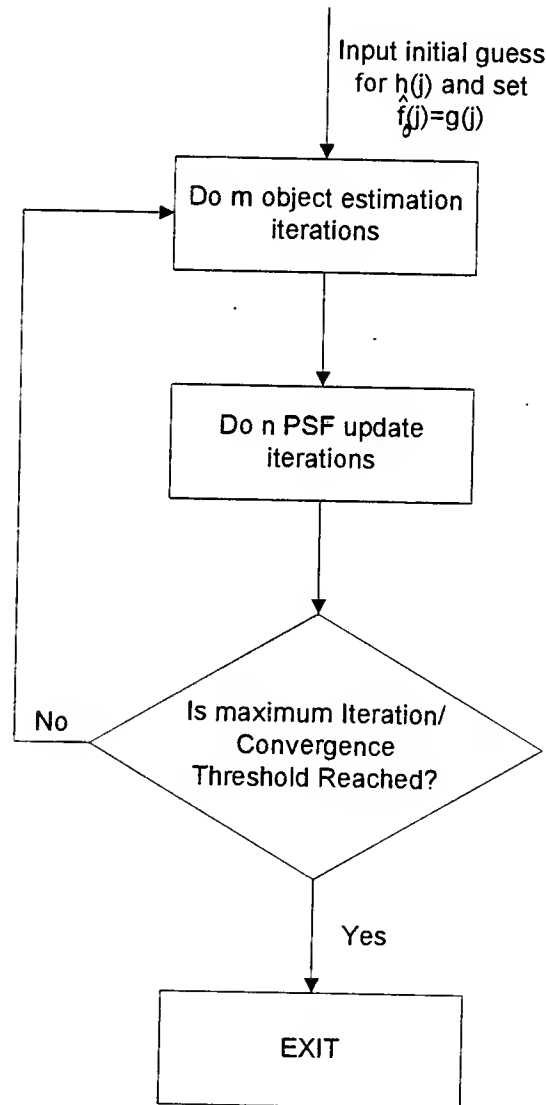


Fig. 12. Flow-chart for implementation of blind ML restoration algorithm.

## 4.2 Performance of Blind ML Restoration Algorithm

Several experiments have been conducted to test the performance of the blind ML restoration algorithm described by (19) and (20) for the superresolution processing of both one-dimensional and two-dimensional signals. A few of these will be described in the following.

### Experiment 1:

Starting with a one-dimensional object consisting of three pulses of uneven heights, shown in Fig. 13a, a blurred image, shown in Fig. 13b, was obtained by convolving with a sensor with cutoff frequency 17 and a OTF whose profile is shown in Fig. 13c. For commencing the blind ML iterations, an initial estimate of PSF was made by assuming a OTF with cutoff frequency 12 (an error was made deliberately to test the performance of the algorithm to yield estimates progressively moving towards the true cutoff frequency of 17) and taking its inverse Fourier transform. The assumed OTF and the PSF are shown in Figs. 13d and 13e.

The OTF estimates resulting after 1 and 3 cycles of the algorithm are shown in Figs. 14a and 14b which indicates the progression of the algorithm in expanding the OTF towards the true cutoff frequency. Each cycle of algorithm implementation consists of five object estimation iterations ( $m = 5$ ) followed by two PSF updating iterations ( $n = 2$ ). The final results after 10 cycles of algorithm implementation are shown in Figs. 14c, 14d and 14e, which show the estimated OTF, the corresponding PSF and the restored object respectively.

### Experiment 2:

In this experiment we tested the performance of the blind ML restoration algorithm in processing PMMW images supplied by the Wright Laboratory Armament Directorate. The specific image used as input to the algorithm is the "Parking Lot 3" image which is shown in Fig. 15a. For obtaining an initial estimate of the PSF to commence the iterations, an analysis similar to the one described earlier of matching the edge profiles was made and an OTF with a cutoff frequency of 82, shown in Fig. 15b (only one dimension of OTF is shown here, for simplicity), was developed. The ML restoration was implemented over 20 cycles with 5 ML estimation iterations ( $m = 5$ ) and 2 PSF updating iterations ( $n = 2$ ) in each cycle. The restored images at the end of 5 cycles and 20 cycles is also shown in Fig. 15e. The progressive enhancement of resolution is clearly evident from the improved structural details of the building and the parked automobiles.

For obtaining a sense of the degree of enhancement of high frequency components, a computation was made to calculate the ratio of the amplitudes of particular spectral components in the processed image to the amplitudes of the same components in the original image. This was conducted by obtaining the 2-dimensional FFT of the original image (in Fig. 15a) and extracting the first row of the Fourier domain amplitude spectrum. The selection of the first row is only for illustrative purposes. By calculating the average of the amplitudes of adjacent 5 pixels, an amplitude histogram, shown in Fig. 16a, was developed. Since the amplitude values in the low frequency range are relatively very high compared to those of the high frequency components, an amplitude thresholding was used to display the high frequency portion more clearly, as shown in Fig. 16b. The corresponding histograms computed from the processed images after 5 and 20 cycles of ML restoration (i.e. from Figs. 15c and 15d) are shown in Figs. 16c and 16d respectively. For giving a better quantitative comparison, the amplitude ratios obtained

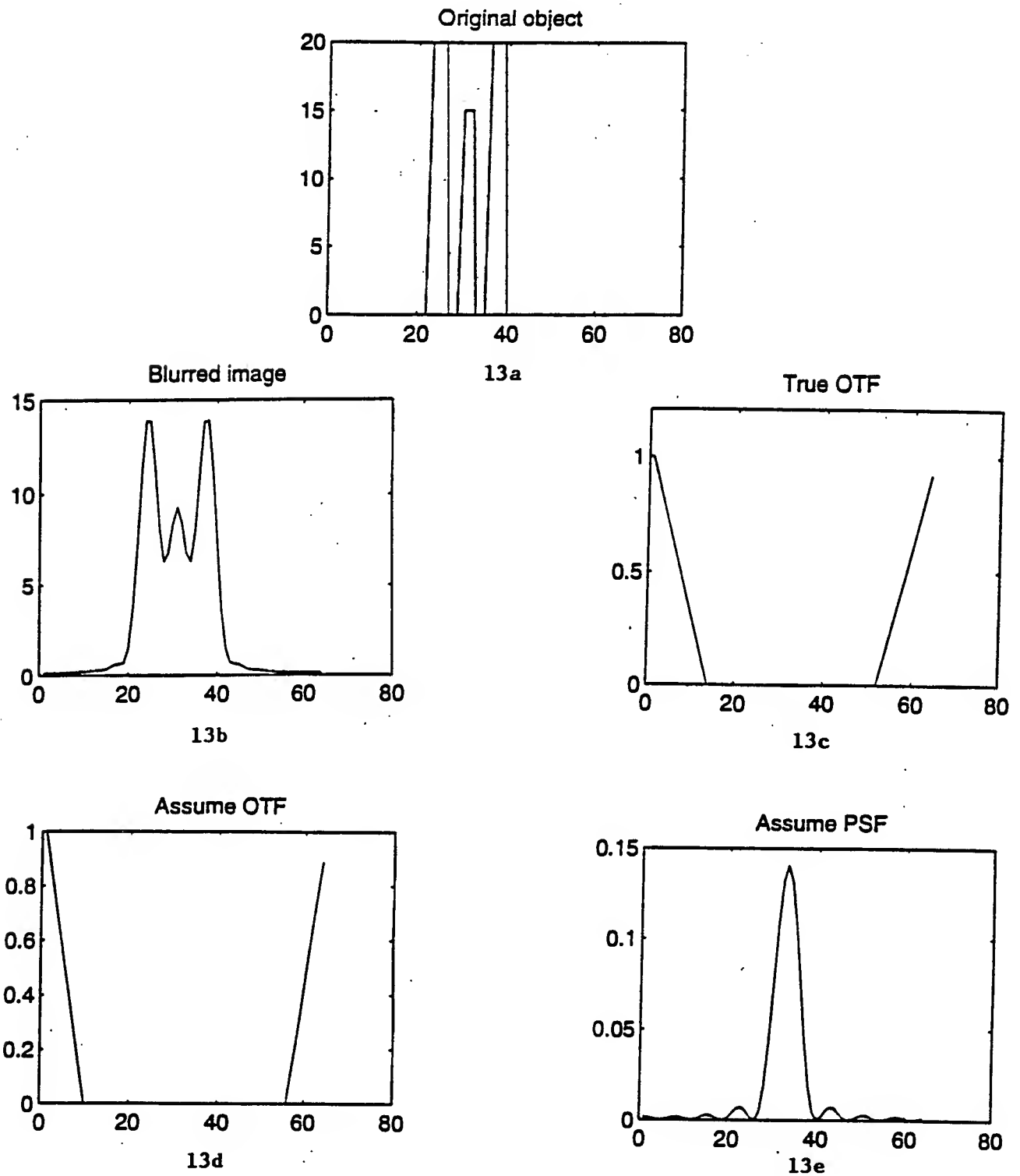


Fig. 13. Signal restoration performance of Blind ML algorithm

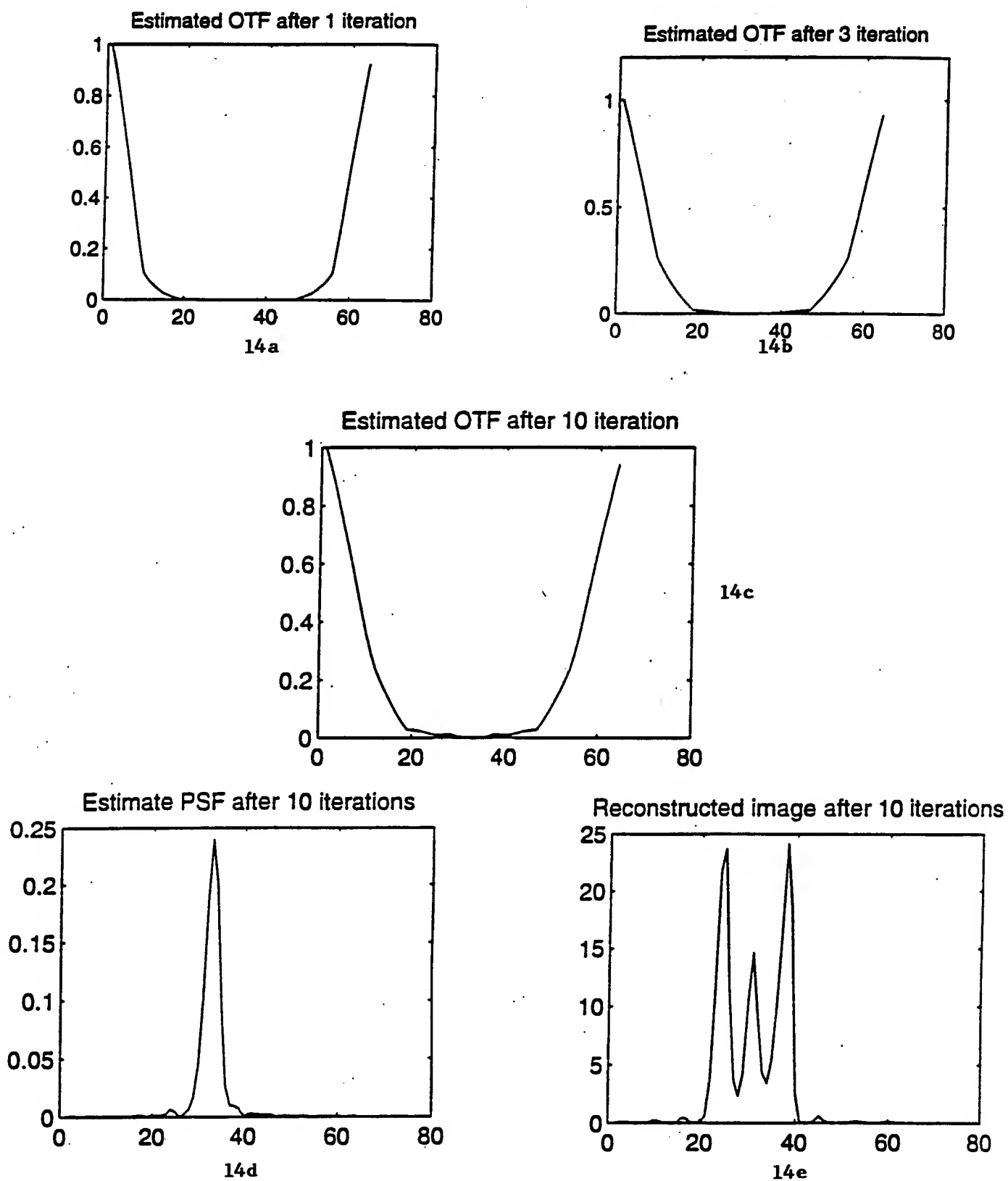
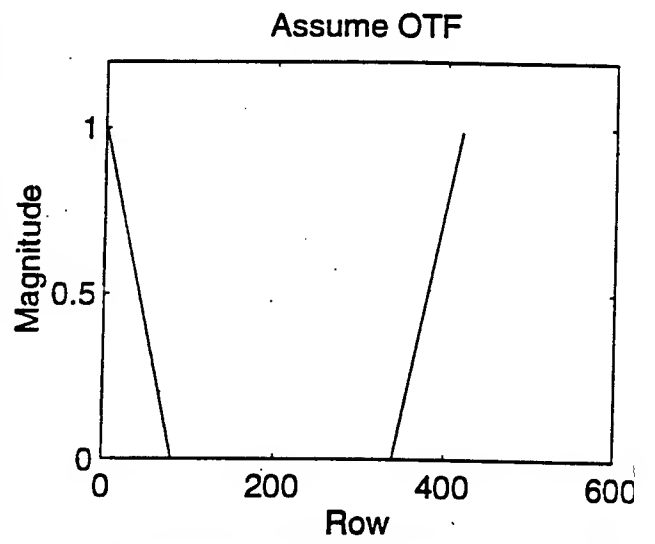


Fig. 14. Blind ML restoration performance in Experiment 1

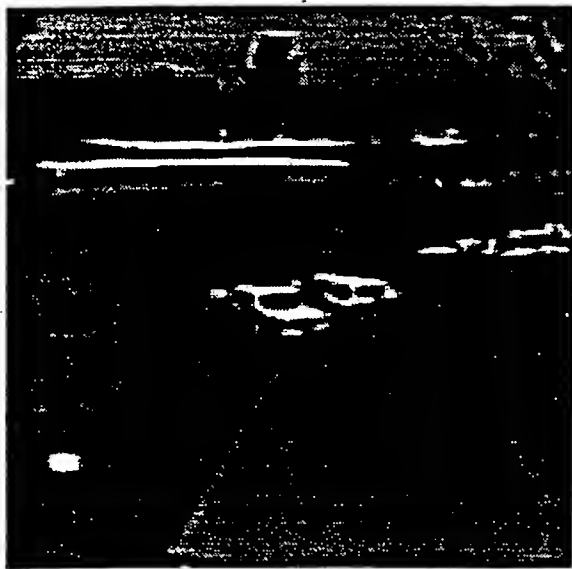




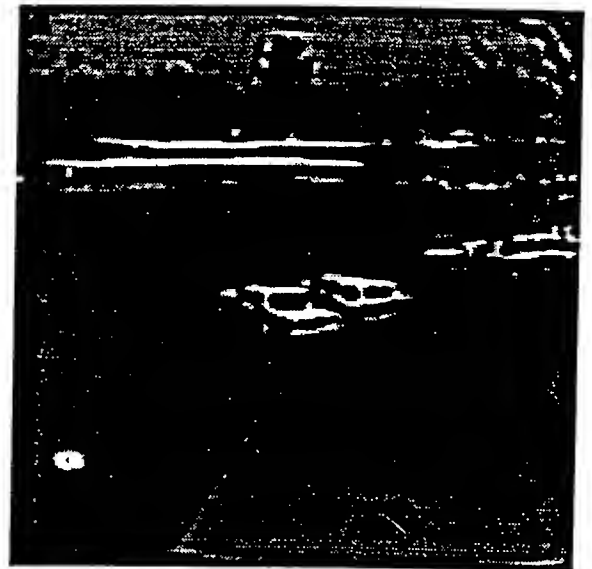
15a



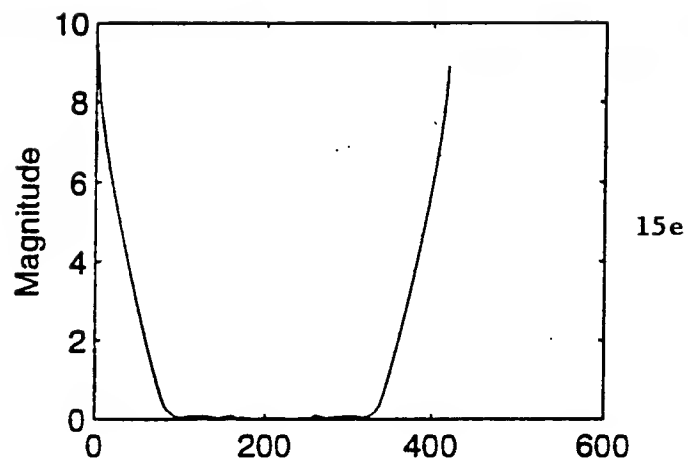
15b



15c



15d



15e

Fig. 15. Blind ML restoration of PMMW image ("Parking Lot 3")...

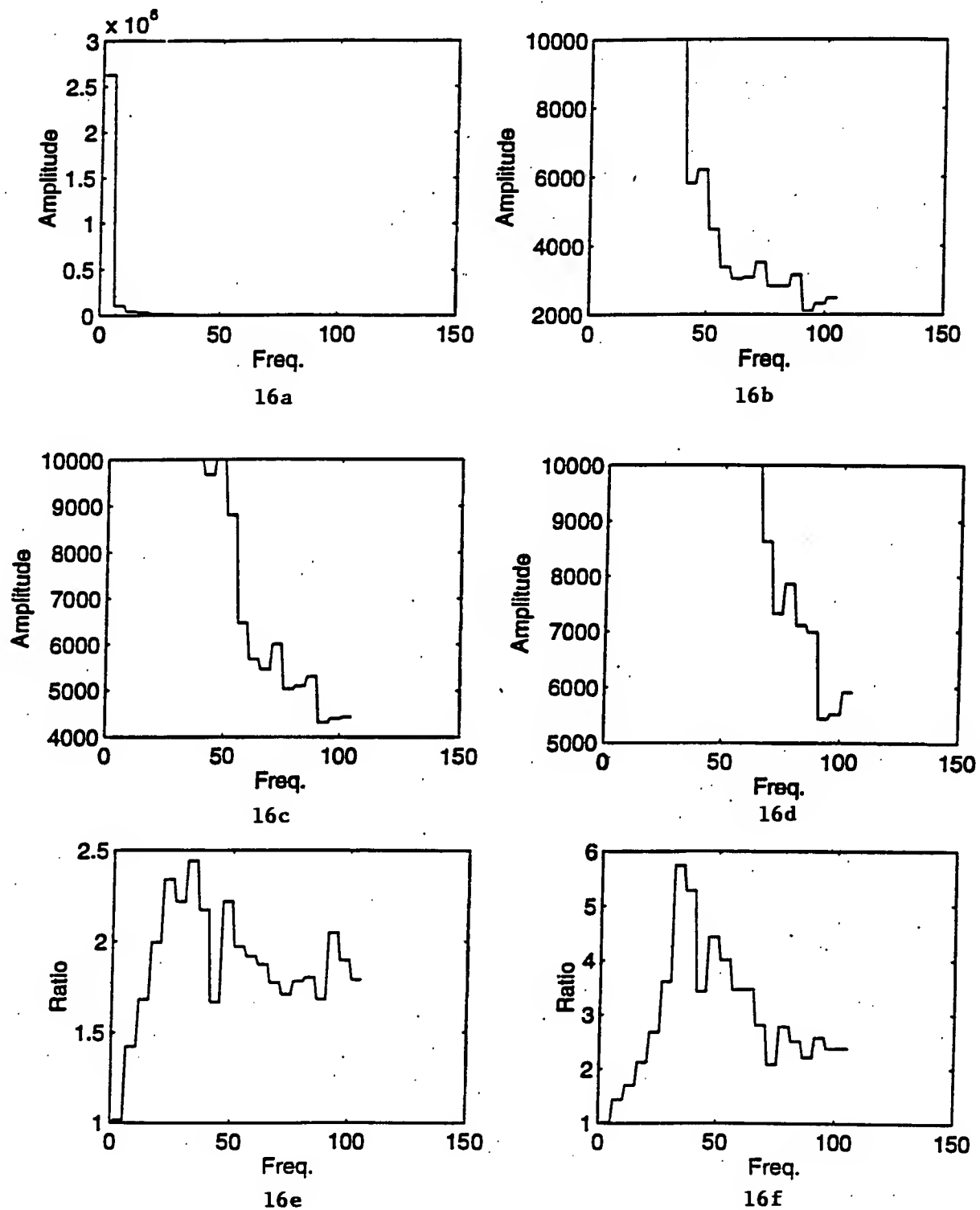


Fig. 16. Frequency enhancement performance of Blind ML algorithm

at the same frequency values were evaluated by computing the amplitude from the processed image divided by the amplitude from the original image. Plots of these amplitude ratios obtained for the images after 5 cycles and 20 cycles of processing are shown in Figs. 16e and 16f, which clearly demonstrate the enhancement of high frequency spectral components by the present ML restoration algorithm.

## 5. CONCLUSIONS

Studies directed to the evaluation of performance of iterative and noniterative schemes for image restoration and superresolution processing were conducted in this project with a specific focus on their eventual deployment in multispectral seeker environments. These studies are of particular relevance to the processing of images collected from various sensors used in these environments to provide enhanced image resolutions that can facilitate better false target rejection, improved automatic target recognition and aimpoint selection. The selection of an iterative technique or a noniterative scheme results in considerably different performance levels and can bring specific advantages and disadvantages to the overall restoration and superresolution function however. Some qualitative comparisons of the performance expected from these schemes were given in this report. A more quantitative performance evaluation of a ML superresolution algorithm was conducted in this project and results of this analysis were also presented here. Finally, to aid in a robust implementation of the ML algorithm in cases where an accurate model of the sensor PSF is not available, a modified algorithm that jointly estimates the object imaged and the PSF parameters from an initial approximated guess of PSF function was developed, and the performance of this blind ML restoration algorithm in processing various signals, including PMMW images supplied by Wright laboratory Armament Directorate Personnel, was described. Based on these studies, it may be concluded that ML techniques offer an attractive framework for tailoring image restoration and superresolution algorithms for implementation in multispectral seeker environments. Further studies directed to improving the computational efficiency of these algorithms and to evaluating the noise tolerance (tradeoff between resolution and noise filtering) are highly useful and these are planned for future investigation.

Acknowledgment: The author is grateful to Mr. Seng-vieng Amphay of the Wright Laboratory Armament Directorate at Eglin AFB for encouragement in this research and for providing PMMW imagery data used in these studies. He also thanks Mr. Ho-Yuen Pang of the University of Arizona Electrical and Computer Engineering Department for his assistance in conducting the investigations contained in this report.

## 6. REFERENCES

1. B.M. Sundstrom and B.W. Belcher, "Smart tactical autonomous guidance", *Proc. of AIAA Missile Sciences Conference*, Feb. 1992.
2. S. Worrell, "Passive millimeter wave imaging sensor assessment", *GACIAC Publication SR-93-03*, Final Report for Wright Laboratory Armament Directorate, Eglin AFB, FL, 1993.

3. M.K. Sundareshan, "Advanced processing techniques for restoration and superresolution of imagery in multispectral seeker environments", *Final Report for Summer Faculty Research Program*, AFOSR, August 1995.
4. W.K. Pratt, "Vector-space formulation of two-dimensional signal processing operations", *Computer Graphics and Image Processing*, Vol. 4, pp. 1-24, 1975.
5. W.K. Pratt, *Digital Image Processing*, Wiley: New York, 1978.
6. A.K. Jain, *Fundamentals of Digital Image Processing*, Prentice-Hall, 1989.
7. C.K. Rushforth and J.L. Harris, "Restoration, Resolution and Noise", *J. Of Optical Society of America*, Vol. 58, pp.539-545, 1968.
8. M.I. Sezan and A.M. Tekalp, "Survey of recent developments in digital image restoration", *Optical Engineering*, Vol. 29, pp.393-404, 1990.
9. J. Biemond, R.L. Lagendijk and R.M. Mersereau, "Iterative methods for image deburring", *Proc. of IEEE*, Vol. 78, pp. 856-883, 1990.
10. D.G. Gleed and A.H. Lettington, "Application of superresolution techniques to passive millimeter-wave images", *Proc. of SPIE Conf. On Applications of Digital Image Processing*, Vol. 1567, pp. 65-72, 1991.
11. R. Bellman, *Matrix Analysis*, McGraw-Hill, 1968.
12. L. Shepp and Y. Vardi, "Maximum likelihood reconstruction in positron emission tomography", *IEEE Trans. On Medical Imaging*, Vol. 1, pp. 113-122, 1982.
13. A.P. Dempster, N.M. Laird and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *J. Of Royal Statistical Society*, Vol. 29, pp. 1-38, 1977.
14. T.L. Ji, M.K. Sundareshan and H. Roehrig, "Adaptive image contrast enhancement based on human visual properties", *IEEE Trans. On Medical Imaging*, Vol. 13, pp. 573-587, 1994.
15. M.K. Sundareshan and F. Amoozegar, "Neural network fusion capabilities for efficient implementation of tracking algorithms", *Proc. SPIE Conf. On Signal and Data Processing of Small Targets, Aerosense' 96*, Orlando, Fl, April 1996. (To appear in expanded from in *Optical engineering*, March 1997).
16. J. Goodman, *Introduction to Fourier Optics*, McGraw-Hill, 1996.

**THE EFFECTS OF METAL INTERLAYER INSERTION ON THE  
FRICTION, WEAR AND ADHESION OF TiC COATINGS**

**Jinke Tang  
Associate Professor  
Department of Physics**

**University of New Orleans  
Lakefront  
New Orleans, LA 70148**

**Final Report for:  
Summer Research Extension Program  
Wright Laboratory**

**Sponsored by:  
Air Force Office of Scientific Research  
Bolling Air Force Base, DC**

**and**

**Wright Laboratory**

**December 1996**

# THE EFFECTS OF METAL INTERLAYER INSERTION ON THE FRICTION, WEAR AND ADHESION OF TiC COATINGS

Jinke Tang  
Associate Professor  
Department of Physics  
University of New Orleans

## Abstract

The friction, wear and adhesion behaviors have been investigated for the TiC coatings containing different kinds of interlayers (Ti, Cr and Mo) of different thickness using pin-on-disk configuration, in which both stainless steel and alumina pins were used. The wear tracks of the TiC coatings and the pins were characterized by both optical microscope or scanning electron microscope (SEM). Inserting Cr or Ti interlayer between TiC coating and substrate could improve greatly the adhesion and wear. The TiC coating with 500Å Cr interlayer showed better overall characteristics than that with 5000Å Cr interlayer. The same was found to be true for the Ti interlayer, *i.e.*, thinner interlayers performed better in the adhesion and wear tests. This is probably due to the improved lattice match between the TiC coatings and the interlayers when the interlayer thickness was reduced. The wear resistance of the TiC coatings was found to increase rapidly as their thickness was changed from 0.2  $\mu\text{m}$  to 2 $\mu\text{m}$ , and to 3  $\mu\text{m}$ , which was closely related to the load bearing capability of the coatings. These TiC coatings with, or without, metal interlayers have been nitrogen ion implanted in the interface region. The effects of such ion implantation on their adhesion and wear are currently being studied.

# THE EFFECTS OF METAL INTERLAYER INSERTION ON THE FRICTION AND WEAR OF TiC COATINGS

Jinke Tang

## Introduction

Because of the many advantages, such as high melting point, high hardness, wear resistance, chemical stability as well as large modulus, ceramic coating materials (e.g., TiC, TiN, SiC, and diamond, *etc.*) have been used in tribological application to reduce the wear of contacting components. The applications range from aerospace, engine parts, to wear resistant barriers for many moving mechanical assemblies [1,2]. Generally, the TiC film deposition has been done by chemical vapor deposition, physical vapor deposition [3,4], magnetron sputtering, pulsed laser deposition (PLD) [5,6], *etc.* Sessler *et al.* [7] investigated the influence of film deposition temperature, substrate hardness, and counterface hardness on the friction and wear behavior of TiC film grown by excimer pulsed laser deposition. The thickness of TiC coatings was controlled at 0.2  $\mu\text{m}$ . Tang *et al.* [8] found the adhesion of magnetron sputtered TiC coating could be modified by inserting metallic interlayer between the coating and stainless steel substrate. The PLD TiC coatings grown at room temperature was harder than the TiC coatings grown by magnetron sputtering under the given experimental conditions.

The purpose of present work was to study the effects of different metallic interlayers (e.g., Cr, Ti, and Mo) on friction and wear of the TiC coating grown by magnetron sputtering. The friction and wear properties of the coatings containing different thickness of TiC film and interlayer were characterized by combining tribological experiments with the optical microscopic analysis.

## Experimental

TiC coatings were grown using the magnetron sputtering technique. The 440C stainless steel (SS) substrates were first rinsed with acetone and isopropanol before they were sputter etched in a diffusion pumped MRC 902 in-line sputter deposition chamber. The chamber was backfilled

with methane and argon at constant flow rate of 40 and 80-90 cm<sup>3</sup>/m, respectively. The bias voltage on substrate was -100V and total pressure was controlled at 8 mtorr. The chambers were pumped to a base pressure of  $1 \times 10^{-6}$  torr before sputter etching of the substrates.

Tribological experiments were carried out on an ISC-200 pin-on-disk tribometer (Implant Sciences Co.) controlled by computer at room temperature in laboratory atmosphere. The sliding speed of pin on disk was constant at 10.21 cm/sec and the loads in all tests were chosen from 25 gram to 500 gram. The substrate, a pellet of 25.5 mm in diameter, was made of 440C stainless steel and the thickness of the TiC coating was 0.2  $\mu$ m, 2  $\mu$ m and 3  $\mu$ m, respectively. Alumina and 440 SS pins with a diameter of 0.125 inch were alternatively used in order to study the influence of ball hardness on friction and wear. Wear traces of the coatings and wear scars of pins were analyzed by optical microscope or scanning electron microscope (SEM) technique.

## Results and Discussion

### 1. Counterface hardness on the friction and wear behavior of the TiC coatings

The friction and wear experiments were conducted using a pin-on-disk configuration, in which the two types of pins, 440 SS and alumina balls, were used. Figure 1 shows the variation of the friction coefficient with the SS ball sliding over the surface of TiC coating with a thickness of 2  $\mu$ m under the loads of 500 g, 250 g, 100 g and 50 g, respectively. It is found in Figure 1 that the friction coefficient increased abruptly, to as high as 0.5-0.9, when the pin started to move on the film. Then it quickly reduced with the sliding, finally reaching an equilibrium value. The value was between 0.2-0.22 for the loads of 50 g, 100 g, and 250 g, but only about 0.12 for 500 g. The initial coefficient of friction in all of above experiments was much higher than the relevant equilibrium coefficient of friction. When the 440 SS ball and the TiC film were brought into contact and sheared, both the adhesion between hard TiC coating and soft SS pin and surface roughness played an important role in determining the initial coefficient of friction. With the SS pin sliding repeatedly over the same track of TiC film, the surface roughness of the pins and TiC film was reduced and the asperities between pin and coating were partly filled by debris (found in the following optical micrograph of the pins). Eventually, a stable equilibrium coefficient of friction was reached. It is noticed easily that the friction coefficient under the load of 500 g was



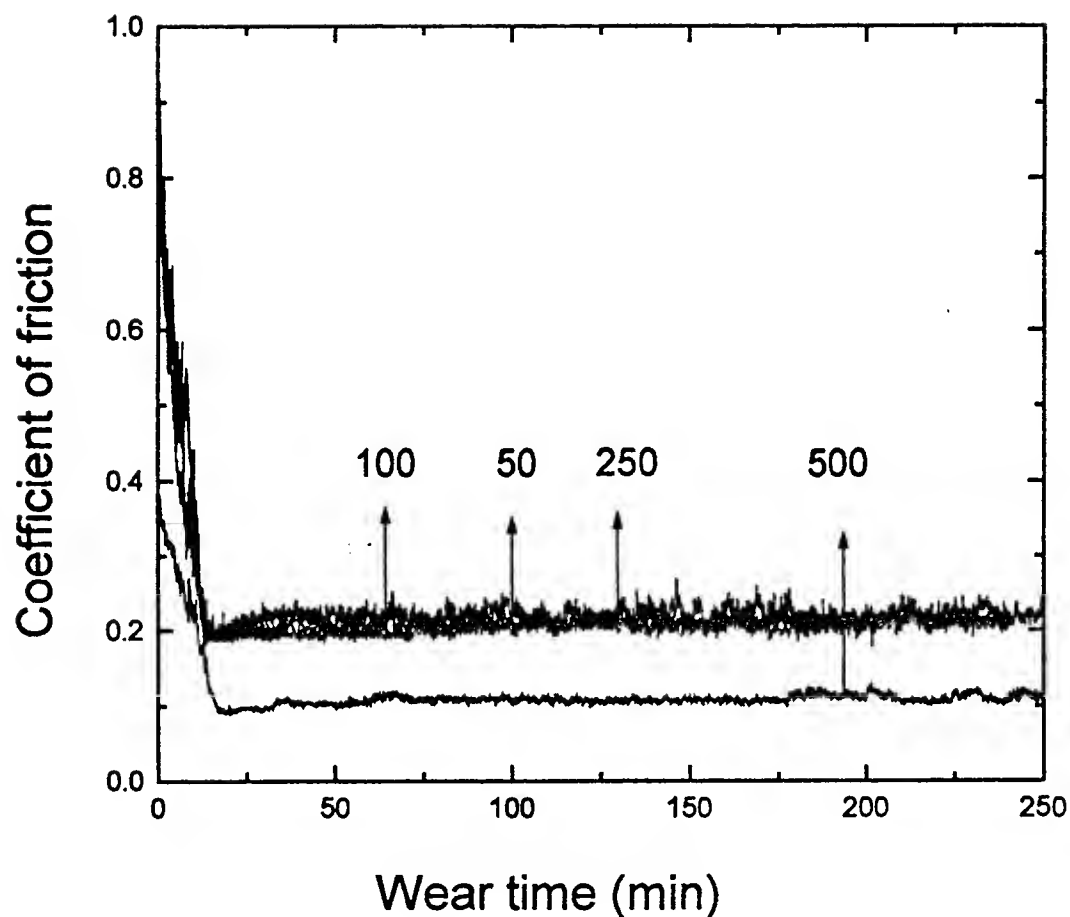


Figure 1 Friction coefficient of the TiC coating(2um) on 440 stainless steel disk as a function of wear time under the loads: 500g, 250 g, 100 g and 50 g, respectively. (speed=10.21 cm/sec, SS pine)

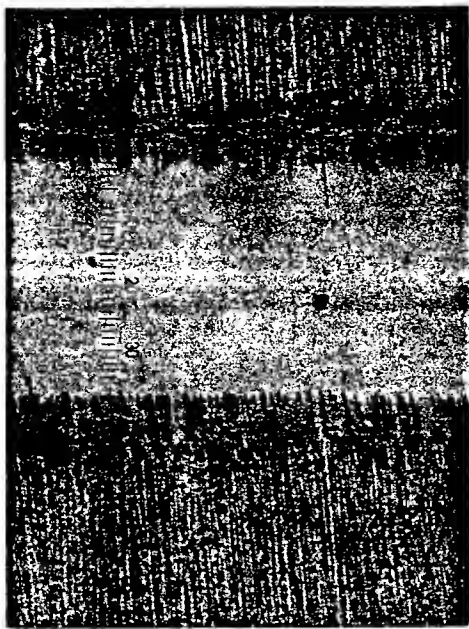
only 0.12, much lower than others. A coefficient of friction of 0.16 was found for a load of 400 g (not shown in Figure 1). This may be due to that the decrease in friction resulting from blunting the tips of surface asperities and reducing surface roughness is greater than the corresponding increase resulting from larger contacting surface when the sliding continued under a heavier load force. Friction tracks of the coating and wear scars of the SS pin shown in the Figure 2 is an evidence of that. Wider wear track on the film and larger wear surface of ball occurred under heavier load, *i.e.*, 500 gram. This load dependence of the friction seems to be directly related to surface roughness and the contacting surface increase.

Friction and wear behavior was investigated by using alumina pin under the same conditions as those using SS pins above. Figure 3 presents the typical friction traces obtained using alumina pin and SS pin in the tests. Both of the equilibrium coefficients of friction were kept in the range of 0.19-0.22. But, the friction coefficient for the TiC/alumina wear couple did not increase so suddenly as that on the TiC/SS wear couple when the pin started to sliding on the surface of TiC film. The initial coefficient of friction of TiC/alumina couple is much lower than that in the case of using SS pin. After sliding of 2500 cycles, the coefficient of friction reached a considerable stable value. The lower initial friction for the TiC/alumina wear couple could probably be related to weak adhesion between harder alumina ball (compared to SS balls) and TiC coating. The friction traces for the TiC/alumina couple in Figure 3(b) exhibited a strong stick-slip characteristic from the beginning of the test, especially after sliding of 23000 cycles. From their micrographs shown in Figure 4, it is found that TiC coating was scratched and apparently worn through by alumina pin, but the contacting surface was only worn a little for the TiC/SS couple. That equilibrium coefficient of friction did not increase to  $\sim 0.6$  (0.6-1.0 for SS substrate) is because the track was not worn through completely. A little scar occurred on the alumina pin, however, there was a large scar on the SS ball. The wear mainly happened on the SS ball and the TiC coating was only worn off to some extent in the experiments using TiC/SS couple.

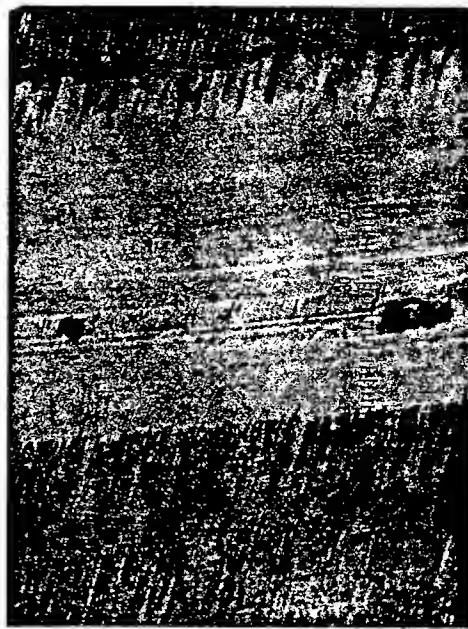
## 2. TiC coating containing metal interlayers

### 2.1. 2 $\mu$ m TiC coating with interlayer

Generally, adhesion between substrate and coating can be greatly affected by interface



(a)



(b)



(c)



(d)

Figure 2 Optical micrographs of worn surfaces of  $2\mu\text{m}$  TiC coating and 440C stainless steel pin under different applied loads. (a)--- 250 g, (b)--- 500 g, (c)--- 250 g, (d)--- 500 g. (Speed=10.21 cm/s)

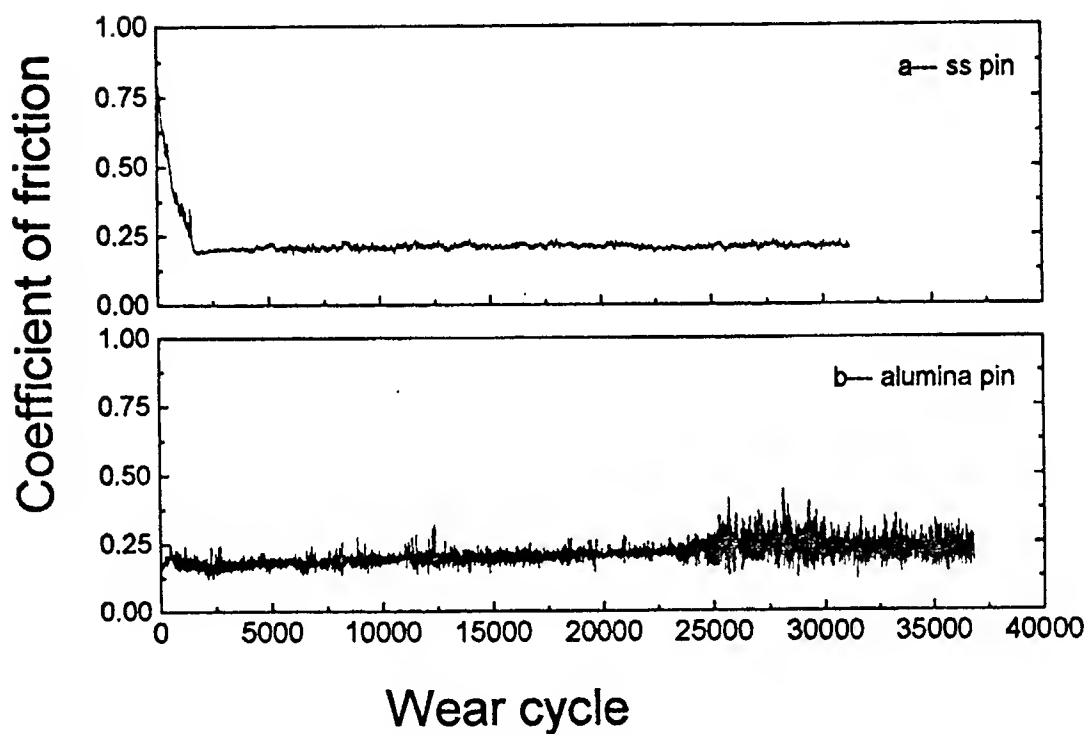
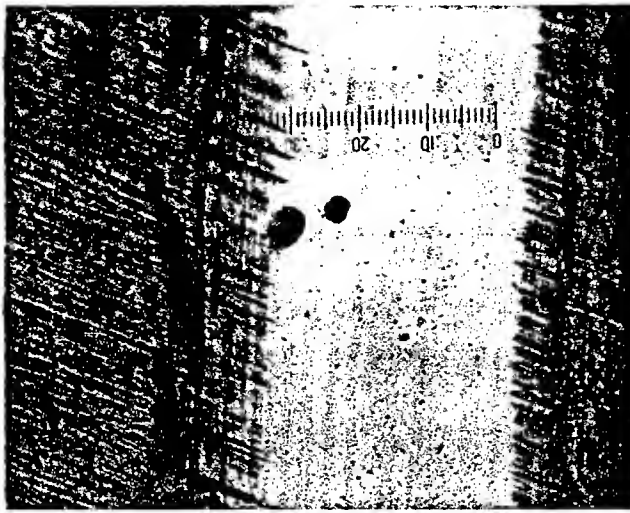
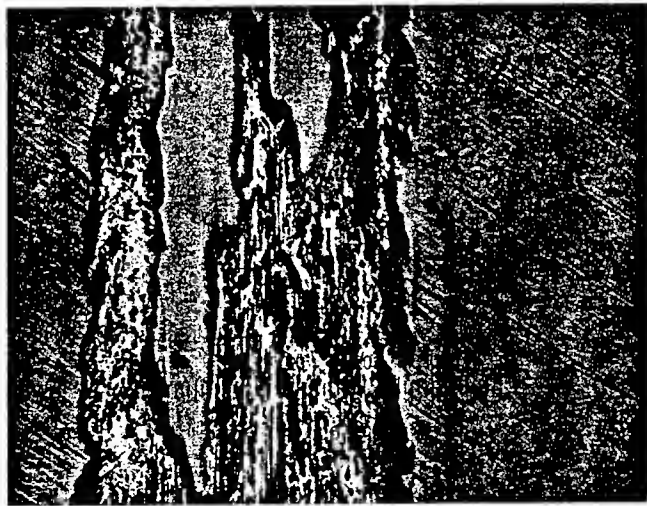


Figure 3 Friction coefficient of the TiC coating on 440 stainless steel disk as a function of wear cycle using both alumina pin and SS pin.( load= 250 g and speed =10.21 cm/sec)



(a)



(b)

Figure 4 Worn traces of  $2\mu\text{m}$  TiC coating under the condition of (a) stainless steel pin and (b) alumina pin. (Load=250g, Time=4 hours, and speed=10.21 cm/s)

structures. Tang and co-workers [8] verified that the adhesion of TiC coating was improved greatly by inserting metallic interlayer, such as Cr, Ti. In this study, the friction and wear has been examined on the TiC coating with interlayers grown by the method of magnetron sputtering, as shown in Figure 5. It is easily seen that the friction curve for the TiC coating containing Cr or Ti interlayer had less stick-slip than that for the TiC coating without interlayer. During the rotation of 23400 cycles (about 3 hours), the sample with 500Å Ti, 500Å Cr or 5000Å Cr interlayer showed more stable friction and wear properties. The friction trace for 5000Å Ti was not as good as the three mentioned above. Equilibrium coefficient of friction for all of the samples kept constant at 0.19-0.22 during the sliding processes.

Their optical micrograph results are presented in Figure 6. Obviously, the TiC coatings were worn off to different degrees and the islands of broken coatings were formed on the wear tracks. Number and size of the islands varied with different samples. The TiC coating containing 500Å Cr, 500Å Ti and 5000Å Cr had less such delamination than that of 5000Å Ti interlayer. The TiC coating with 500Å Cr interlayer showed the best resistance to wear in all of the samples. This was due to a very strong adhesion, and maybe hardness, after Cr interlayer was inserted between the substrate and coating [8]. The photo results in Figure 6 also illustrated that friction and wear behavior was better for the coating containing the thinner (500Å) Cr interlayer than that with thicker (5000Å) Cr interlayer. Although the coatings showed good friction properties after Ti interlayer was inserted, it was not very effective when the inserted Ti interlayer was as thick as 5000Å. The largely destroyed coating on the wear trace in the photograph of TiC coating with 5000Å Ti interlayer was consistent with the large stick-slip phenomenon seen in the friction curve. On the wear tracks of the TiC coating containing interlayers, there were significant streak lines, especially around the islands. It seems that, after alumina pin was slid over the surface of TiC coating for some times, the cracks appeared gradually along some certain directions on the wear surface. The cracks were caused by the ball sliding repeatedly on the coating surface under loads. It is interesting that the streak line often appeared on the surface of TiC coating containing interlayer Ti or Cr under load of 100-250 g (Figure 7). In contrast, there was no apparent streak line for coatings without an interlayer, even after the load was increased up to 500 g, at which the coating had been seriously delaminated. It seems that the TiC coating without interlayer was

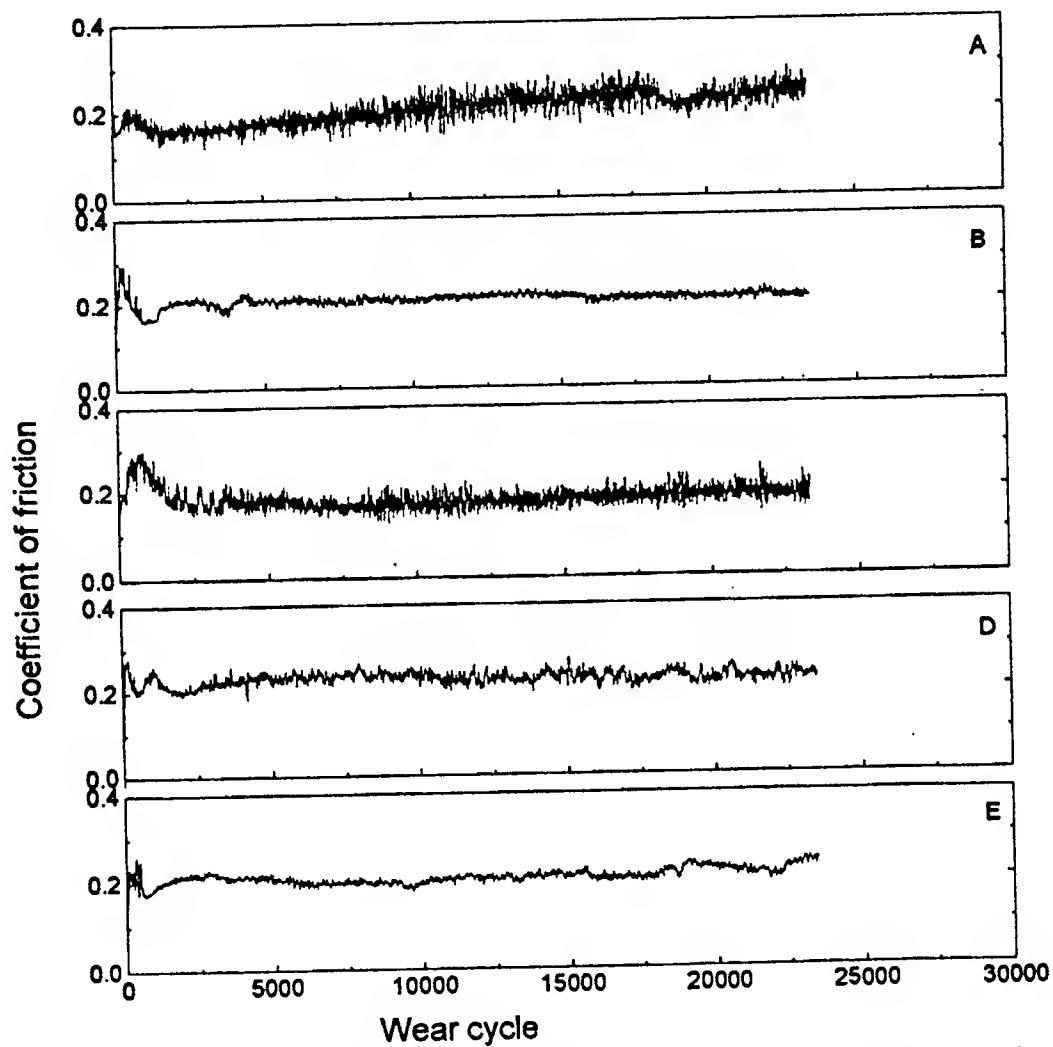
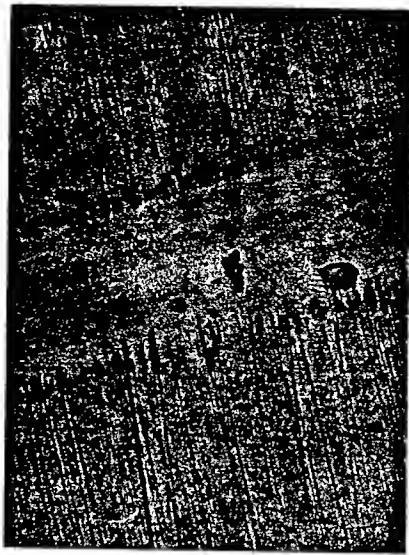
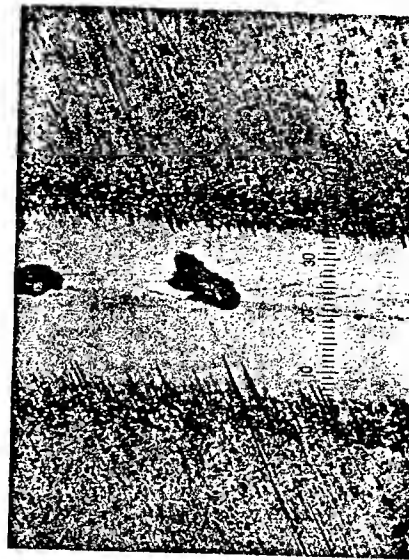


Figure 5 Friction coefficient of the TiC coating(2 $\mu$ m) without and with interlayers on stainless steel substrate as a function of number of passes under the load 250 g and speed 10.21cm/s. A, no interlayer; B, 500A Ti; C, 5000A Ti; D, 500A Cr; E, 5000A Cr.



(a)



(b)



(c)



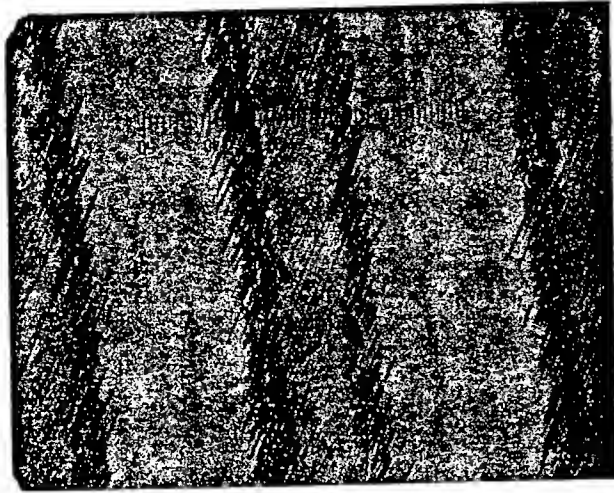
(d)



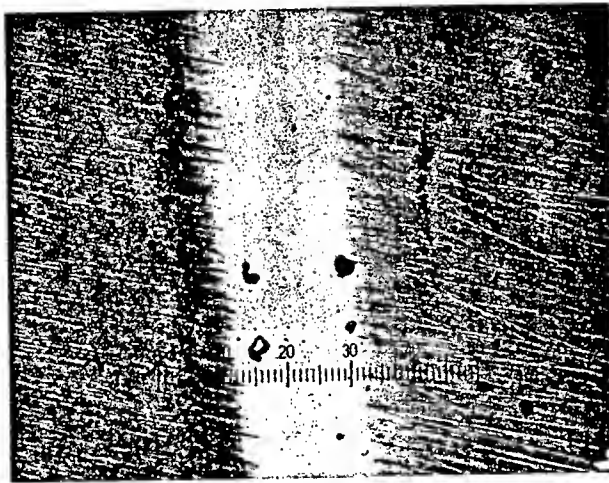
(e)

Figure 6 Optical micrographs of the worn traces for  $2\mu\text{m}$  TiC coating with different interlayers, (a)---  $500\text{ \AA}$  Cr, (b)---  $5000\text{ \AA}$  Cr, (c)---  $500\text{ \AA}$  Ti, (d)---  $5000\text{ \AA}$  Ti, (e)--- no interlayer. (Load=250 g, Time=3 hours, and Alumina pin).





(a)



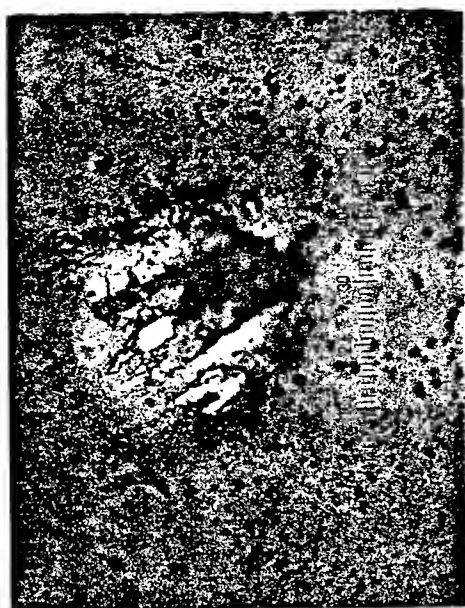
(b)

Figure 7 Worn traces of  $2\mu\text{m}$  TiC coating containing (a)  $500\text{ \AA}$  Ti interlayer and (b) without interlayer. (a)--- 100 g load (left) and 150 g load (right), (b)--- 100 g load.

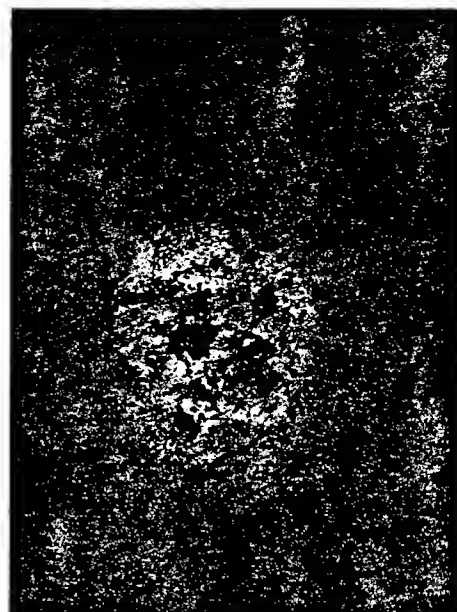
directly delaminated because of the weak adhesion between TiC coating and SS substrate, while coatings containing Ti or Cr interlayer first underwent a process of wear track cracking before the delamination happened on the contacting surface. This wear track cracking might be attributed to the existence of softer Ti or Cr interlayers, but the enhanced adhesion due to the interlayers was able to prevent substantial delamination. The cracks on wear track should be ominous of the islands for the TiC coating with Ti or Cr interlayer. There were fewer cracked lines on friction surface of TiC coating with 500Å Cr interlayer. The surface of TiC coating with 5000Å Ti interlayer was seriously scratched after 3 hours of wear. On the other hand, the alumina pin was also abraded while the TiC coating was worn through. It can be seen in Figure 8 that a large ball scar was formed as the TiC coating was scratched seriously, especially for TiC coating with 5000Å Ti interlayer.

Figure 9 is the coefficients of friction of TiC coatings in long distance tests of over 3500 meters (about 10 hours). The tests were done to further examine the friction and wear life of the coatings. The variation of curves is similar to that of Figure 5. But, the friction for the sample containing 5000Å Ti interlayer and without interlayer varied unstably in the later stage of sliding. Friction coefficient for TiC coating containing 500Å and 5000Å Cr interlayers started to increase after sliding of ~3400 meters. Combining with the analysis of optical micrograph of the TiC coating, as shown in Figure 10, we found that the wear track of TiC coatings was scratched and delaminated greatly in the experiments, especially for the coating without and with 5000Å Ti interlayers, and extended broken coating was formed instead of islands. Although there were heavy scratch on the surface of TiC coating with 500Å Ti, 500Å Cr and 5000Å Cr interlayers, their friction coefficients varied still in the range of 0.2-0.3 during the sliding processes. This indicates that even though part of TiC coating was scratched off in the friction test, some TiC debris probably filled in the gap between the contacting surface of ball and TiC coating to some extent such that the friction coefficient increased only slightly. For heavy delamination of the of the TiC film, the friction coefficient was changed significantly, for example, in the case of TiC coating without or with 5000Å Ti interlayer. Similar to the results in Figure 5 and 6, the long scratch on the wear surface was accompanied with large scars on the ball surface.

The behavior of friction and wear of the TiC coatings could be improved apparently after



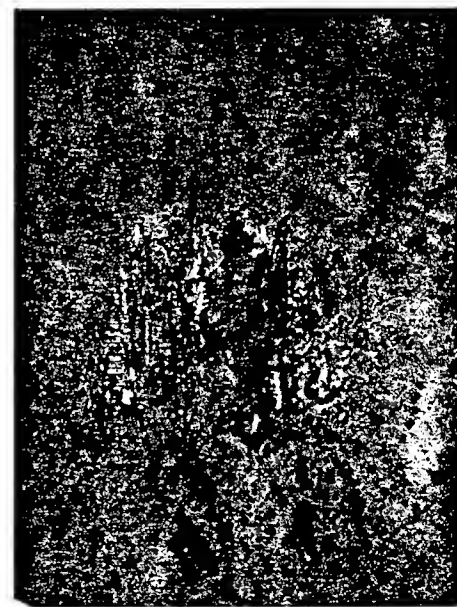
(a)



(b)



(c)



(d)

Figure 8 Optical micrographs of the alumina pin scar coupled with TiC coating containing different interlayers. (a)--- 500 Å Cr, (b)--- 5000 Å Cr, (c)--- 500 Å Ti, (d)--- 5000 Å Ti.

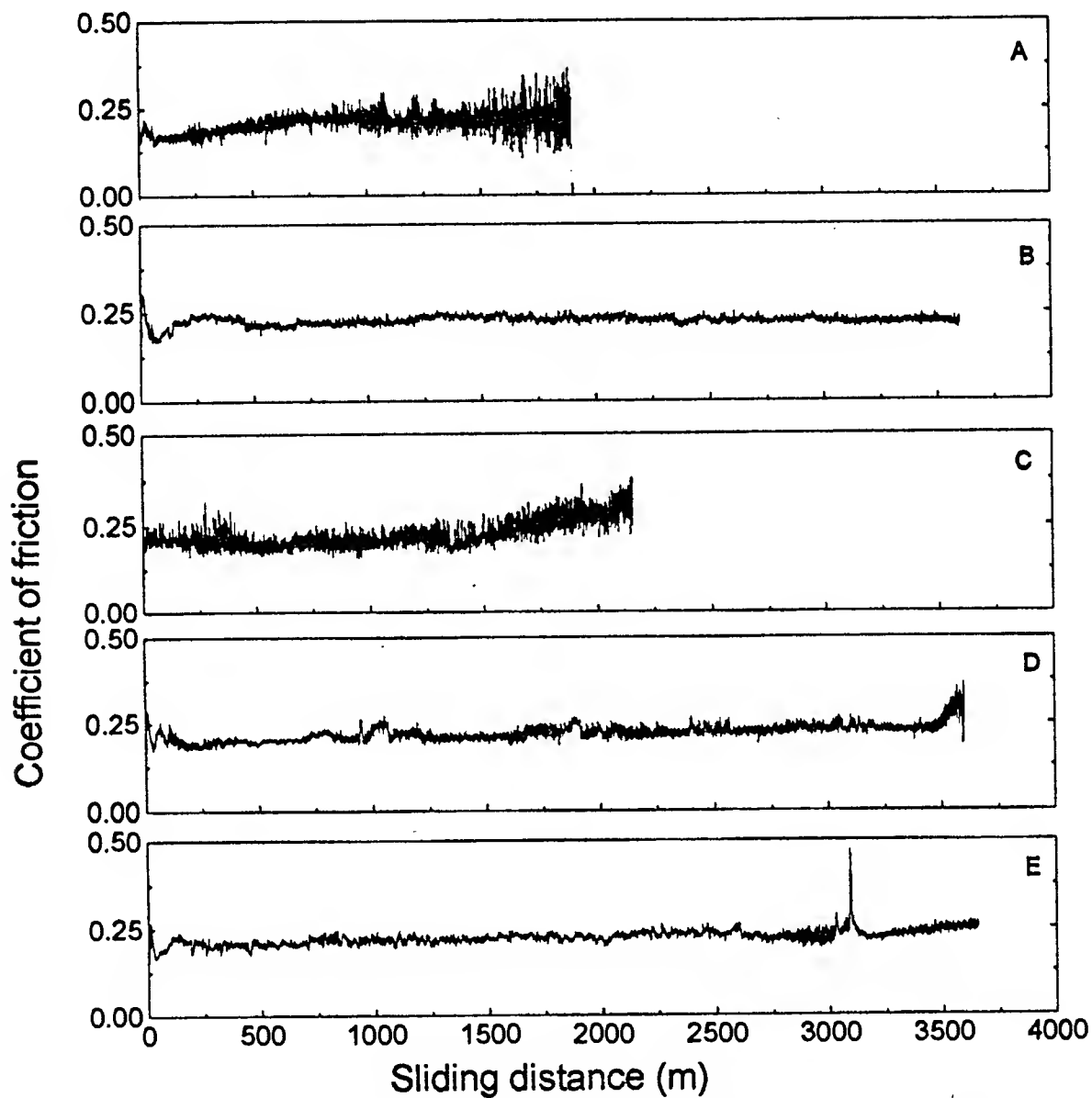


Figure 9 Friction coefficient of the TiC coating (2μm) with and without interlayers as a function of sliding distance. A, no interlayer; B, 500A Ti; C, 5000A Ti; D, 500A Cr; E, 5000A Cr. ( Load=250g and speed=10.21 cm/sec).



(a)



(b)



(c)



(d)



(e)

Figure 10 Optical micrographs of the  $2\mu\text{m}$  TiC coating with different interlayers after a wear time of 9.8 hours, (a)---  $500\text{ \AA}$  Cr, (b)---  $5000\text{ \AA}$  Cr, (c)---  $500\text{ \AA}$  Ti, (d)---  $5000\text{ \AA}$  Ti, (e)--- no interlayer. (load=250 g, speed=10.21 cm/s, and alumina pin).

Ti or Cr element was inserted between TiC films and stainless steel substrates, although the effects on the tribological properties were different for each kind of the interlayer. The microstructure of the coatings have been characterized with scanning electron microscopy (SEM) technique. Figure 11 shows the cross-section features of the 2  $\mu\text{m}$  TiC coatings with the interlayer 500Å Cr, 1000Å Cr, 5000Å Cr, 500Å Ti, 1000Å Ti, and 5000Å Ti, respectively. The pictures were magnified by 13,000. The interfaces between the substrate and interlayer Cr or Ti, and between the interlayer and TiC film can be seen clearly in the figure. It is known that the TiC is harder than the substrate and interlayer Cr or Ti. Inserting a layer of soft Ti or Cr between the TiC film and substrate should be able to dissipate the stress due to applied load, which could enhance the adhesion of the TiC coatings. Compact interfaces between coating, interlayer, and substrate were seen except the sample with 5000Å Ti interlayer. The interface between the TiC coating and the 5000Å Ti interlayer is not as good as that between the Ti interlayer and the substrate. This interface quality should have direct effects on the adhesion the coatings. The friction and wear results discussed earlier reflected this and were consistent with the interface structural analysis.

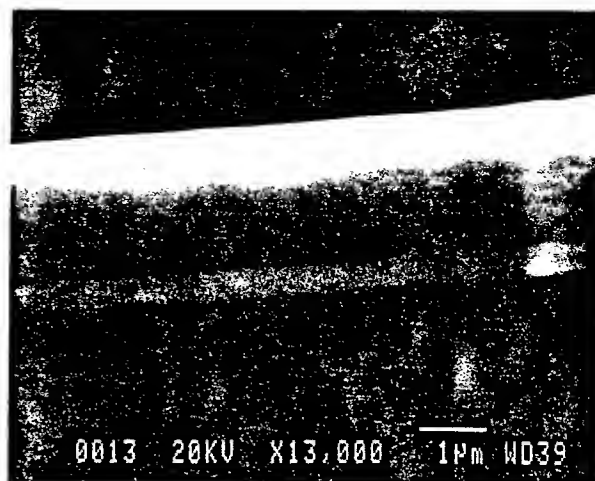
Table 1 lists the volume loss of alumina and stainless steel pins sliding over 2.0 $\mu\text{m}$  TiC coating with and without interlayer. Clearly, The rates of volume loss for the worn SS spherical pin is at least twice as large as that for the alumina pin. In contrast, the results have also proved that TiC coating is easier to be worn off using alumina pin than using SS pin. There was no apparent difference among the TiC coatings with different interlayers and without interlayer as alumina pin was used in friction experiments. It seems that the volume loss of SS spherical pin is increased with the increasing of the applied load, especially, there was a large difference in the rate between 500 g and 50 g.

## 2.2. 3 $\mu\text{m}$ TiC coating

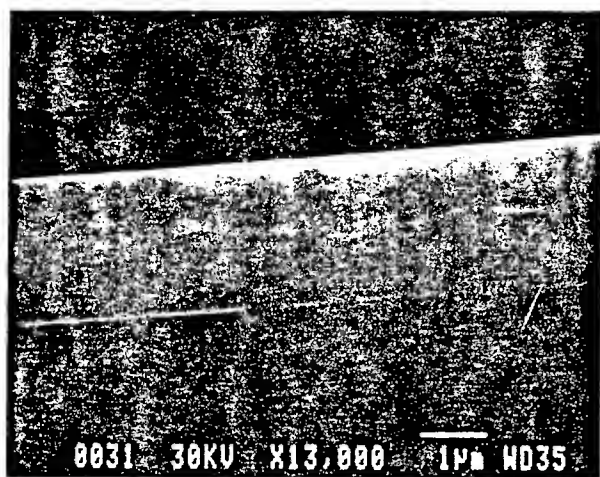
Increase of the thickness of TiC coating should also have a influence on the friction and wear of the coating. Figure 12 shows the friction coefficient of 3 $\mu\text{m}$  TiC coatings without interlayer and with interlayers, 1000Å Cr, 1000Å Ti, and 1000Å Mo, respectively, as a function of the number of sliding passes. It is obvious that friction for the coatings with Cr, Ti, as well as



( a )



( b )



( c )



( d )

Figure 11 SEM images of the cross-section of  $2\mu\text{m}$  TiC coating with different interlayers, (a)---  $500\text{ \AA}$  Cr, (b)---  $5000\text{ \AA}$  Cr, (c)---  $500\text{ \AA}$  Ti, (d)---  $5000\text{ \AA}$  Ti.

Tabl 1 Volume loss of worn spherical pins coupled with TiC coatings. (a)--- stainless steel pin, (b)---alumina pin

(a)

Load (gram)	500	250	100	50
Worn Rate of Pin (mm <sup>3</sup> /H)	$2.23 \times 10^{-4}$	$8.00 \times 10^{-5}$	$7.99 \times 10^{-5}$	$3.67 \times 10^{-5}$

(b)

Coating Components	2 $\mu\text{m}$ TiC +5000 Å Ti	2 $\mu\text{m}$ TiC +500 Å Ti	2 $\mu\text{m}$ TiC +5000 Å Cr	2 $\mu\text{m}$ TiC +500 Å Cr	2 $\mu\text{m}$ TiC
Worn Rate (mm <sup>3</sup> /H)	$3.00 \times 10^{-5}$	$2.97 \times 10^{-5}$	$2.99 \times 10^{-5}$	$2.98 \times 10^{-5}$	$3.02 \times 10^{-5}$



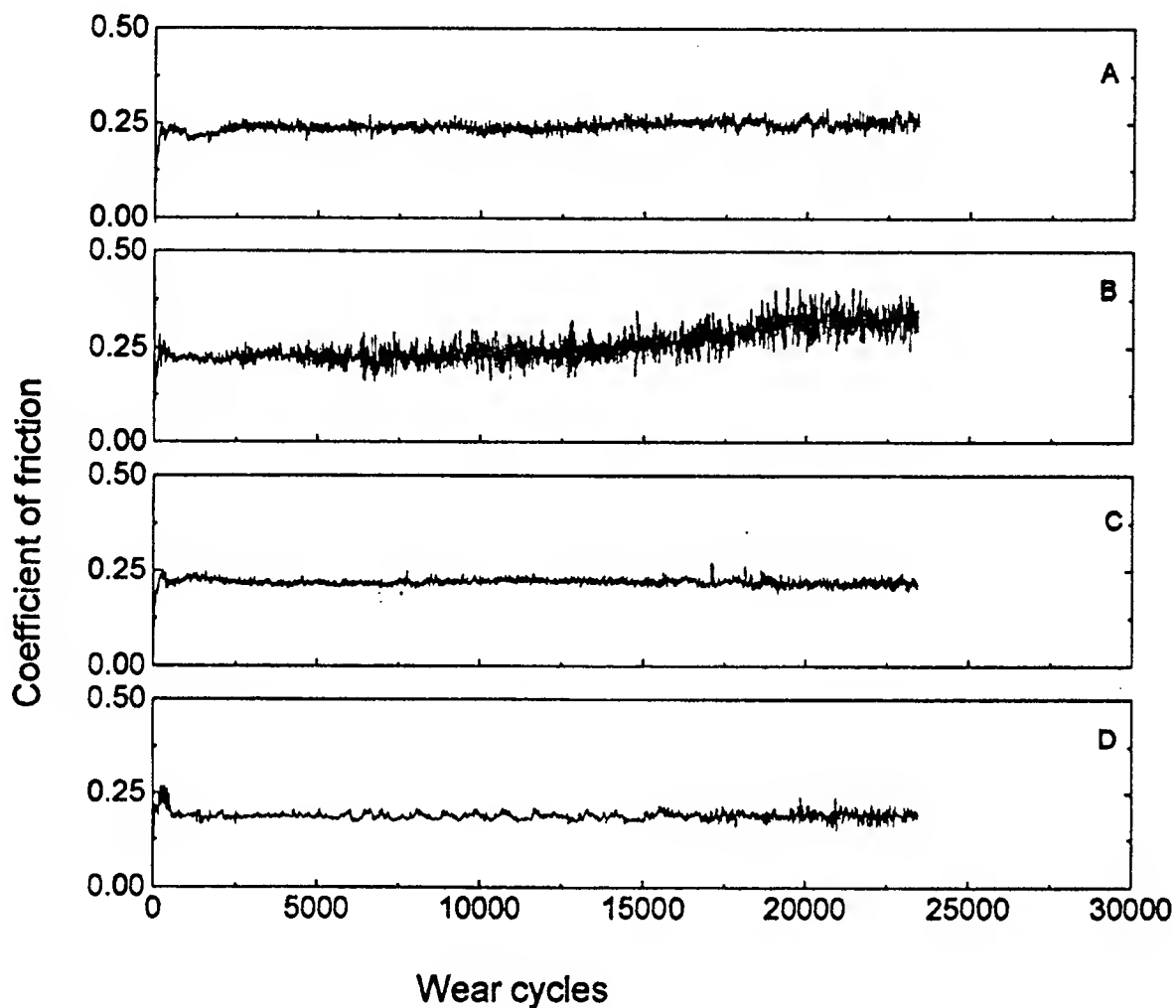


Figure 12 Friction coefficient of the TiC coatings(3um) with and without interlayer as a function of number of passes under the load 250 g and speed 10.21 cm/sec. A, no interlayer; B, 1000Å Mo; C, 1000Å Ti; D, 1000Å Cr.

without interlayers were more stable and smooth than that containing Mo interlayer during 23400 cycles of wear. This is probably because of very poor adhesion of the coating containing Mo interlayer, so that the TiC coating could not adhere to the substrate strongly. The scratch occurred easily on the TiC coating surface, resulting in increased friction (Figure 12(b)). The optical micrographs in Figure 13 indicated that the damage on the surface of the TiC coating with Cr interlayer was the smallest in all of the wear tests, only showing a beginning of the wearing off. In contrast, TiC coating containing Mo was greatly delaminated. The TiC coating with Ti interlayer showed nearly the same as that without interlayer in friction and wear, but the former had smoother friction curve than the latter. When the sliding distance was extended to over 3500 meters, it is concluded further from Figure 14 that TiC coating with Cr (or Ti) interlayer has a better wear life than the TiC coating without an interlayer. The coefficient of friction of the latter varied over a large range and increased rapidly in the later stage of wear test. In fact, the optical micrograph results gave a directly evidence for such observation (Figure 15).

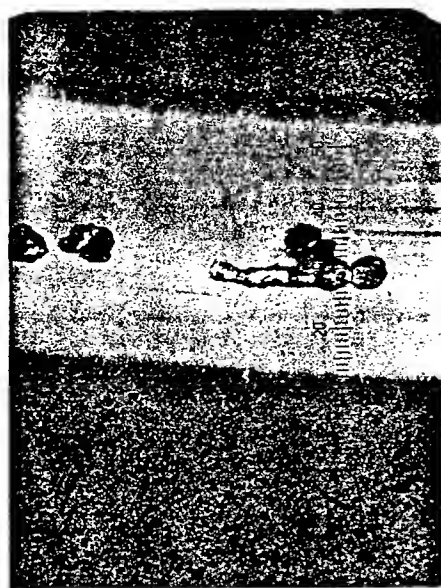
Considering the influence of the thickness of TiC coatings on adhesion and wear, we found that the TiC coating of  $3\mu\text{m}$  thick was much better than the TiC coating of  $2\mu\text{m}$  thick.

### 2.3. $0.2\mu\text{m}$ TiC coating

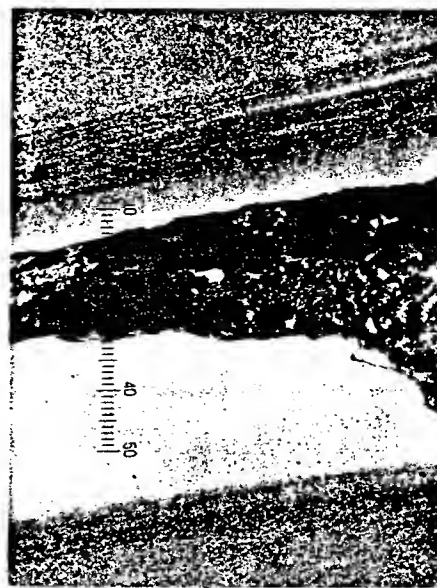
The friction and wear characteristics of the coatings should be affected by applied load, especially when TiC coating is very thin. Figure 16 and 17 demonstrate the friction as a functions of the sliding time for the thin TiC coatings ( $0.2\mu\text{m}$  in thickness) containing  $200\text{\AA}$  Ti and  $200\text{\AA}$  Cr interlayers, respectively, under different loads. The coefficient of friction of the TiC coating with Ti interlayer in Figure 16 was stable during the sliding of 3 hours when the load was below 50 g. With increasing load, the friction and wear of the TiC coating became gradually unstable, especially when load exceeded 100 g, due to the severe delamination of the coating. Similar to Figure 16, the friction and wear behavior for the TiC coating with Cr interlayer in Figure 17 was better under a load of 50 g, compared to a load of 100 g.

## Summary and Conclusions

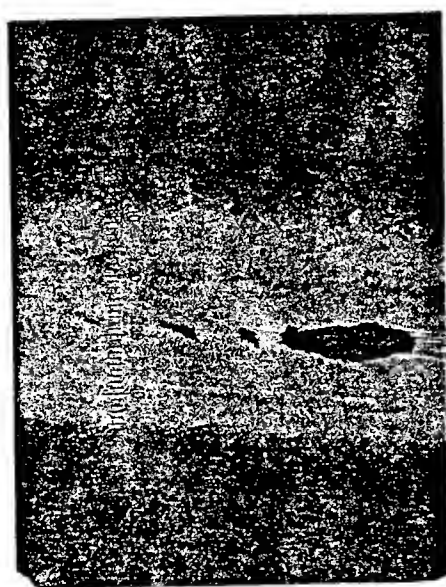
The friction, adhesion and wear behaviors have been investigated for the TiC coatings



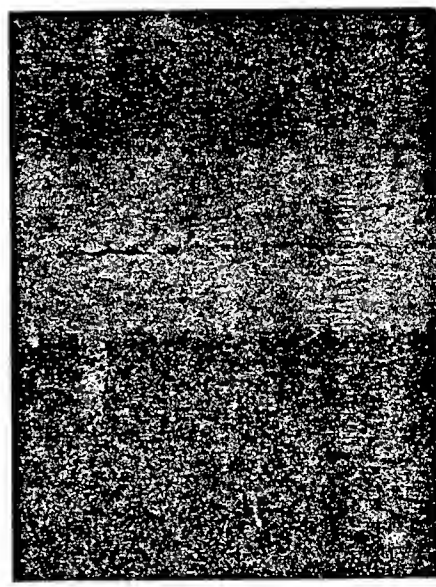
(a)



(b)



(c)



(d)

Figure 13 Optical micrographs of the worn traces of  $3\mu\text{m}$  TiC coating with different interlayers, (a)--- no interlayer, (b)---  $1000\text{ \AA}$  Mo, (c)---  $1000\text{ \AA}$  Ti, (d)---  $1000\text{ \AA}$  Cr. (Load=250 g, time=3 hours, and alumina pin).

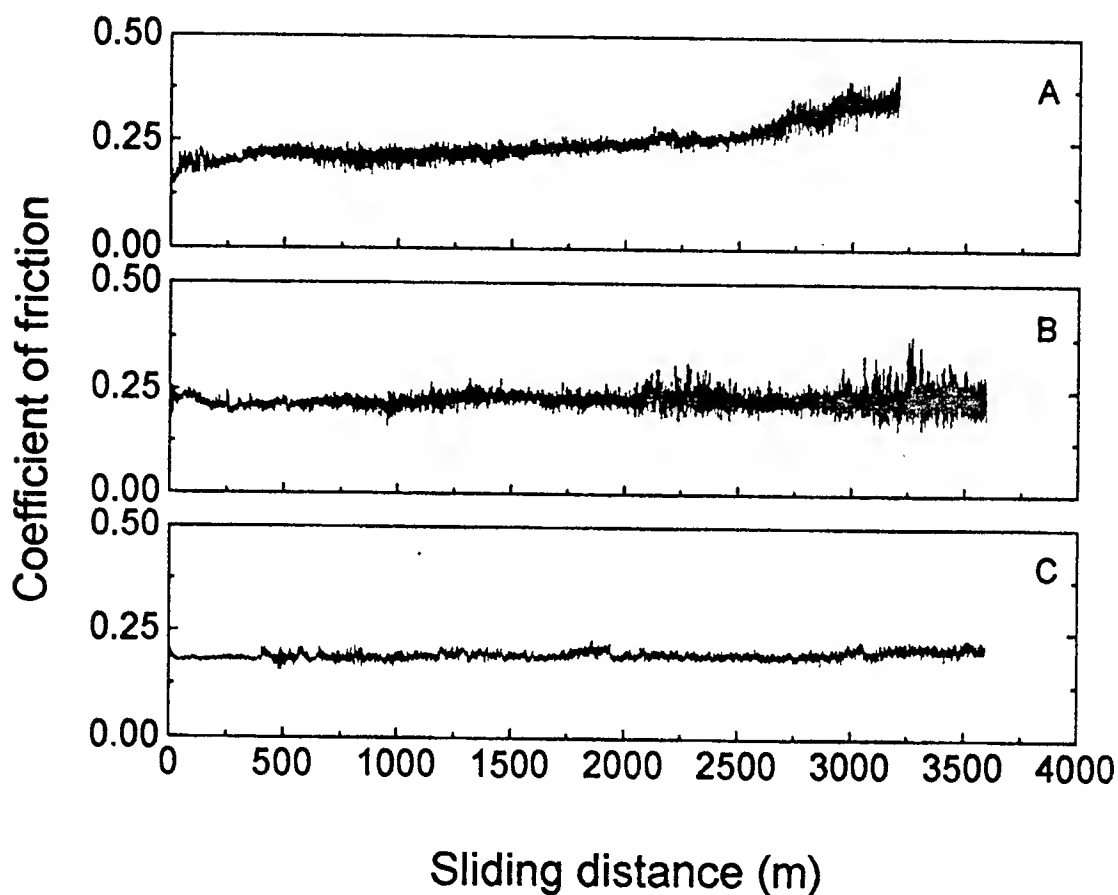
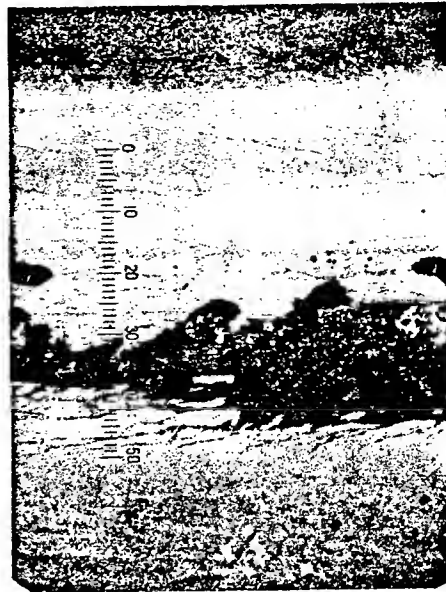


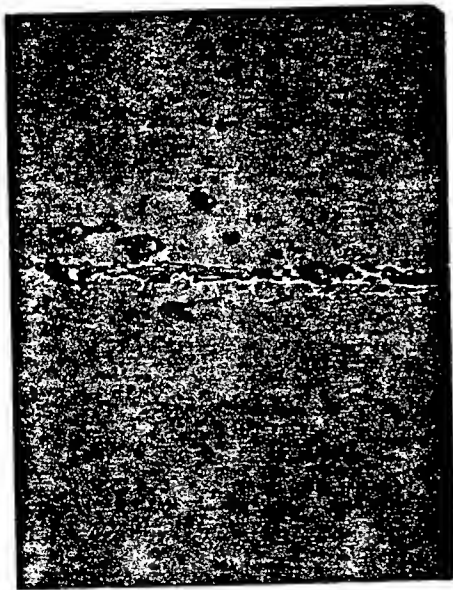
Figure 14 Variation of the friction coefficient of TiC coatings (3 $\mu$ m) with and without interlayer with sliding distance. A, no interlayer; B, 1000Å Ti; C, 1000Å Cr. (Load=250 g; sliding speed=10.21cm/sec).



(a)



(b)



(c)

Figure 15 Worn traces of the  $3\mu\text{m}$  TiC coating with different interlayers after a worn time of 9 hours. (a)--- no interlayer, (b)---  $1000\text{ \AA}$  Ti, (C)---  $1000\text{ \AA}$  Cr. (Load=250 g, speed= 10.21 cm/s, and alumina pin).

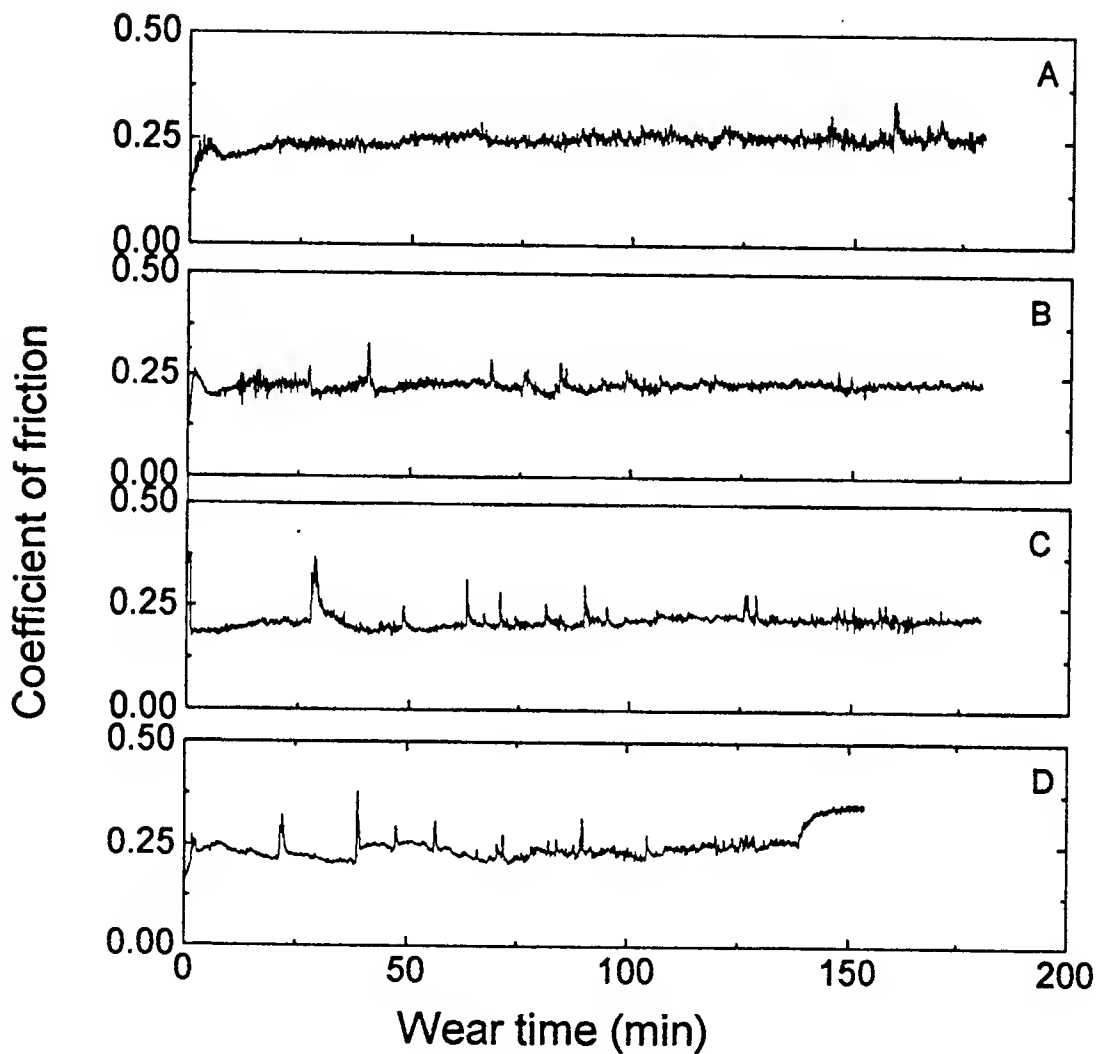


Figure 16 Friction coefficient of the TiC coating(0.2 $\mu$ m) with 200 Å Ti interlayer as a function of wear time under the different loads, A--- 25 g, B--- 50 g, C--- 100 g, and D--- 150 g. (speed=10.21 cm/sec).

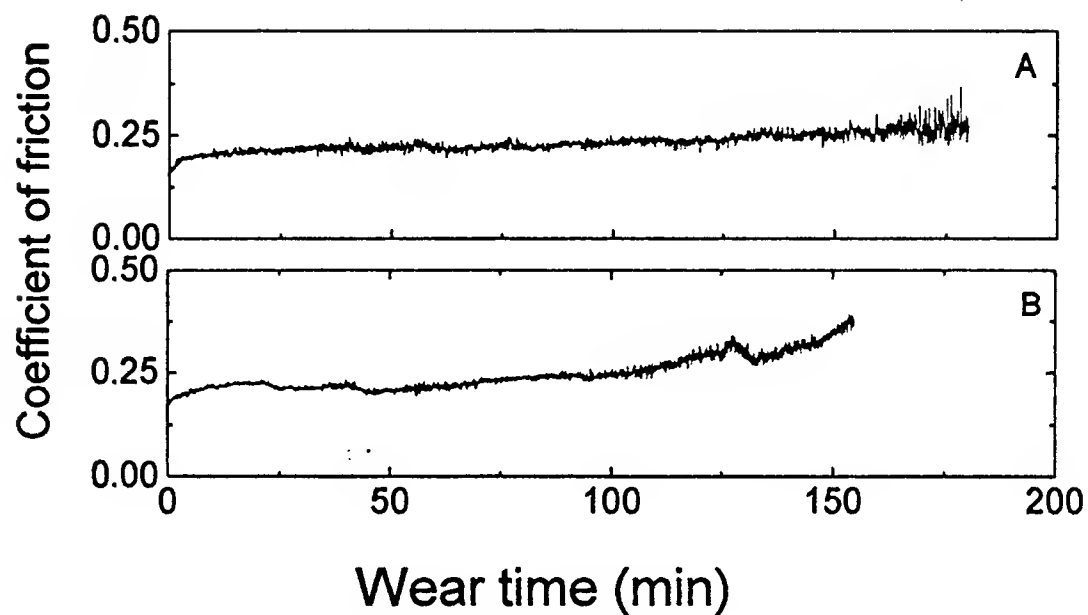


Figure 17 Friction coefficient of the TiC coating(0.2 $\mu$ m) with 200Å Cr interlayer as a function of wear time under the load. A--- 50 g, and B--- 100 g. speed=10.21 cm/s.

containing different kinds of interlayers (Ti, Cr and Mo) of different thickness using pin-on-disk configuration, in which both stainless steel and alumina pins were used. The wear tracks of the TiC coatings and the pins were characterized by both optical microscope or scanning electron microscope (SEM).

- 1) The equilibrium coefficients of friction were found to vary between 0.19 and 0.22 for the TiC coatings as the stainless steel and alumina pins were alternatively used in our experiments. There was significant stick-slip during the wear tests for TiC/alumina couple. According to their optical microscopic results, it was found that TiC coating was scratched and apparently worn through by alumina pins, but only slightly by stainless steel pins. The scars on the alumina pins were much less severe than the scars on SS pins during the wear test. The rate of volume loss of SS pins was more than twice of that of alumina pins.
- 2) Inserting Cr or Ti interlayer between 2  $\mu\text{m}$  TiC coating and substrate can improved greatly the adhesion and wear. It is found that the TiC coatings containing 500Å Cr, 5000Å Cr, 500Å Ti and 5000Å Ti interlayers showed more stable and smoother friction than that without interlayer during 3 hours of sliding test. The microscopic analyses further indicated the coating with 500Å Cr, 500Å Ti, 5000Å Cr and 5000Å Ti interlayers had stronger adhesion and better wear than the coating without interlayer. The TiC coating with 500Å Cr interlayer performed better than that of 5000Å Cr interlayer in friction and wear tests. The same was found to be true for the Ti interlayer, *i.e.*, thinner interlayers performed better in the adhesion and wear tests. This is probably due to the improved lattice match between the TiC coatings and the interlayers when the interlayer thickness was reduced.
- 3) When the friction test was extended to about 10 hours long, that is, the sliding distance of over 3500 meters, the TiC coating with 500Å Cr, 500Å Ti and 5000Å Cr interlayers maintained good friction and wear behavior, but the TiC coating with 5000Å Ti or without interlayer did not. The wear surface of the latter has been destroyed much more seriously than that of the former.
- 4) The wear resistance of the TiC coatings could be increased by changing their thickness from 0.2  $\mu\text{m}$  to 2  $\mu\text{m}$ , and to 3  $\mu\text{m}$ . This was closely related to the load bearing capability of coatings of different thickness. The friction and wear behaviors of the thin TiC coating was easily affected by applied load. For load of 150 g or above, the 0.2 $\mu\text{m}$  TiC coating with 0.02 $\mu\text{m}$  Cr or Ti



interlayer could be destroyed by alumina pins in a short sliding time.

5) During 3 hours (about 23400 cycles) of sliding tests of 3  $\mu\text{m}$ -thick TiC coatings with 1000Å Cr, 1000Å Ti and 1000Å Mo interlayers, it is found that TiC coatings with Cr or Ti interlayers have much stable and smooth friction and wear behavior, compared to that containing Mo interlayer. This is because of the poor stress bearing capability of the porous Mo film deposited at low temperature. The degree of delamination on the surface of TiC coating with Cr interlayer was the smallest in all of the samples.

## References

1. J. C. Angus and C. C. Hayman, Science 241, 913(1988).
2. L. Kempfer, Mater. Eng. 108, 28(1991).
3. G.Georgiev, N. Feschiev, D. Popov and Z. Uzuuov, Vacuum, 36(1986)595.
4. J. F.Sundgren, B. O. Johansson and S. E. Karlsson, Thin Solid Films, 105(1983)353.
5. O. Rist and P. T. Murray, Mater. Lett., 10(1991)322.
6. M. S. Donley, J. S. Zubinski, W. J. Sessler, V. F. Dyhouse, S. D. Walck and N. T. McDevit, Matter. Res. Soc. Symp. Proc. 236(1991)461.
7. W.J. Sessler, M. S. Donley, J. S. Zabinski, S. D. Walck and V. J. Dyhouse, Surface and Coatings Technology, 56(1993)125-130.
8. Jinke Tang, Jeffrey S. Zabinski and J. E. Bultman, Surface and Coatings Technology, (1997) In press.

**DEVELOPMENT OF MASSIVELY PARALLEL EPIC HYDROCODE  
IN CRAY T3D USING PVM**

**C.T. Tsai  
Associate Professor  
Department of Mechanical Engineering**

**Florida Atlantic University  
777 Glades Road  
Boca Raton, FL 33431**

**Final Report for:  
Summer Faculty Research Program  
Wright Laboratory**

**Sponsored by:  
Air Force Office of Scientific Research  
Bolling Air Force Base, DC**

**and**

**Wright Laboratory**

**December 1996**

# DEVELOPMENT OF MASSIVELY PARALLEL EPIC HYDROCODE IN CRAY T3D USING PVM

C.T. Tsai  
Associate Professor  
Department of Mechanical Engineering  
Florida Atlantic University

## Abstract

The objective of this report is to verify the feasibility of converting a large sequential EPIC hydrocode into a massively parallel EPIC hydrocode in terms of computational speed. Sequential subroutines in the Research EPIC hydrocode, a Lagrangian finite element analysis code for high velocity elastic-plastic impact problems, are individually converted into parallel code using Cray Adaptive Fortran (CRAFT). The performance of massively parallel subroutines running on 32 PEs on Cray-T3D is faster than their sequential counterparts on Cray-YMP. At next stage of the research, Parallel Virtual Machine (PVM) directives is used to develop a PVM version of the EPIC hydrocode by connecting the converted parallel subroutines running on multiple PEs of T3D to the sequential part of the code running on single PE. With an incremental increase in the massively parallel subroutines into the PVM EPIC hydrocode, the performance with respect to speedup of the code increased accordingly. The results indicate that significant speedup can be achieved in the EPIC hydrocode when most or all of the subroutines are massively parallelized.

# **DEVELOPMENT OF MASSIVELY PARALLEL EPIC HYDROCODE IN CRAY T3D USING PVM**

C. T. Tsai

## **1. INTRODUCTION**

The aim of research and development of weapon systems is to develop one which can respond quickly to the combat needs of operational commanders. Presently, warhead design involves expensive and time consuming tests done on prototypes. The computer simulation of warhead penetration can save huge amounts of money and time on warhead design. The simulation of complex penetration behavior requires large amount of CPU time, even when run on fast vector processor machine like the Cray-YMP. The EPIC hydro code is used, at the Armament Directorate of Air Force Wright Laboratory, to solve high velocity elastic-plastic impact problems involved in warhead design.

The primary objective of the thesis was to verify the feasibility of converting a large finite element analysis code into a massive parallel finite element code in terms of computational speed and cost. The Research EPIC hydro code , a Lagrangian finite element analysis code, was selected for parallization. A computational intensive algorithm, a large data set and a liberal data distribution policy was a motivation for parallelizing the EPIC code. The Cray-T3D was chosen as the MPP platform as it supports MIMD/SPMD model, CRAFT programming model and PVM message passing programming model. More importantly, it is closely coupled with other Cray PVP systems like Cray-YMP on which the EPIC code is already developed. Related goals included developing an incremental approach of parallization using PVM message passing paradigms and identifying CRAFT parallization techniques best suited for the EPIC subroutines.

The next section gives a brief introduction to parallel processing and its various applications. Chapter 2 describes the EPIC theory. Cray-T3D architecture and MPP programming model is discussed in Chapter 3. Chapter 4 states the performance results of the individual parallel subroutines. Chapter 5 describes the incremental approach and porting of EPIC subroutines using PVM. Chapter 6 contains the implementation of EPIC hydro code using heterogeneous PVM. Chapter 7 include conclusions and future work.

## **1.1 INTRODUCTION TO PARALLEL PROCESSING**

The ever increasing computational needs of emerging applications have been the primary motivating factor for the steady increase in speed of the traditional serial computers. However, fundamental physical limitation imposed by the speed of light makes it impossible to achieve further improvements in speed of serial or single processor computers infinitely. Recent trend show that the performance of these computers is beginning to saturate. ( It is often remarked that speeds of basic microprocessor grow by a factor of 2 every 18 months; this empirical observation, true over many years, is called Moore's law.). A natural way to circumvent this saturation is to use an ensemble of processors to solve problems [1].

Parallel processing, the method of having many small tasks solve one large problem, has emerged as a key enabling technology in modern computing. The past several years have witnessed an increasing acceptance and adoption of parallel processing, for both high performance scientific computing and for more general purpose applications, was a result of the demand for higher performance, lower cost and sustained productivity. The acceptance of parallel processing has facilitated two major developments: massive parallel processors(MPPs) and the use of distributed computing.

A massively parallel processing (MPP) machine combine a few hundred to a few thousand CPUs in a single large cabinet connected to hundreds of Gbytes of memory. MPPs offer enormous computational power and are considered to be one of the most powerful computers in the world. MPPs are used to solve computational Grand challenge

problems such as global climate modeling, determining molecular, atomic and nuclear structures. As simulations become more realistic, the computational power required to produce them grows rapidly and that is when MPPs come into picture.

The second major development affecting scientific problem solving is distributed computing. Distributed computing is a process whereby a set of computers connected by a network are used collectively to solve a single large problem. As more and more organizations have high speed local area networks interconnecting many general purpose workstations, the combined computational resources may exceed the power of a single high performance computer. In some cases, several MPPs have been combined using distributed computing to produce unequalled computational power [2].

## **1.2 MOTIVATION AND APPLICATIONS FOR PARALLEL COMPUTING**

The traditional scientific paradigm is first to do theory, and then lab experiments to confirm or deny the theory. The traditional engineering paradigm is first to do a design and then build a laboratory prototype. Both paradigms are being replaced by numerical experiments and numerical prototyping for the following reasons: real phenomena are too complicated to model on paper (e.g. climate prediction) and real experiments are too hard and too expensive for a laboratory (e.g. oil reservoir simulation, large wind tunnel, overall aircraft design etc.).

Scientific and engineering problems requiring the most computing power to simulate are commonly called "Grand Challenges", like predicting the climate few years hence, are estimated to require computers computing at the rate of 1 Tera flops (i.e.,  $10^{12}$  floating point operations per second), and a memory size of 1 TB (Tera Byte). One of the "Grand Challenge" climate modeling problem is illustrated. In a simplified climate model, climate is defined as function of 4 arguments: longitude, latitude, elevation and time. This in turn, returns a vector of 6 values: temperature, pressure, humidity, and wind velocity (3 variables actually). To represent the continuous function in the computer, the domain is discretized and climate is evaluated for the arguments lying on a grid: climate (i, j, k, n),

where  $t = n \cdot dt$ , where  $dt$  is a fixed time step,  $n$  an integer, and  $i, j, k$  are integers indexing the longitude, latitude and elevation grid cells, respectively.

An algorithm to predict the weather (short term) or climate (long term), is a function which maps the climate at time  $t$ ,  $\text{Climate}(i, j, k, n)$ , for all  $i, j, k$ , to the climate at the next time step  $t+dt$ ,  $\text{climate}(i, j, k, n+1)$ , for all  $i, j, k$ . The algorithm involves a system of equations including, in particular the Navier-Stokes equations for the fluid flow of gases in the atmosphere. Then, the earth's surface is discretized into 1 kilometer by 1 kilometer cells in the latitude -longitude direction, and 10 cells in the vertical direction. From the surface area of the earth, we can compute that there are about  $5 \cdot 10^9$  cells in the atmosphere. With six 4-byte words per cell, the memory requirement is about 0.1 TB.

Assuming, it takes 100 flops to update each cell by one minute. Or in other words, if  $dt = 1$  minute, and computing  $\text{climate}(i, j, k, n+1)$  for all  $i, j, k$  from  $\text{climate}(i, j, k, n)$  take about  $100 \cdot 5 \cdot 10^9$  or  $5 \cdot 10^{11}$  floating point operations. For such kind of computations, the computing speed has to be atleast 8 Gflops. Weather prediction (computing 24 hours to compute the weather 7 days hence), requires computing  $50 \cdot 12 = 600$  times faster, or 4.8 Tflops machine.

The actual grid resolution used in climate modeling today is about 4 degrees latitude by 5 degrees of longitude (about 460 km by 560 km), a rather coarse resolution. A near term goal is to decrease this grid size using parallel computing techniques, such that different grid data are stored in different processors and relevant simultaneous equations are solved parallelly on different processors [3].

Selected application areas in the field of parallel computing are as follows:

- Weather and Climate
  - ◊ Prediction of weather, climate and global change
- Biology



- ◇ Determination of molecular, atomic and nuclear structure.
- ◇ Mapping the human genome and understanding the structure of biological macromolecules.
- Mechanical Engineering
  - ◇ Finite element Methods.
  - ◇ Particle methods in aerospace.
  - ◇ Understanding turbulence, pollution dispersion and combustion system.
  - ◇ Computational Fluid Dynamics.
- Chemical Engineering
  - ◇ Gas hydrate simulation.
  - ◇ Simulation of fluids in pores.
- Environmental modeling
  - ◇ Modeling whole ecosystems.
  - ◇ Assessment of pollution remediation.
- Material science
  - ◇ Understanding the nature of new materials.
  - ◇ Exploring theories of matter.
- Demography
  - ◇ Interactive access to large databases.

### **1.3 CLASSIFICATION OF PARALLEL COMPUTERS**

Parallel computers are classified based on various dimensions like control mechanism, address-space organization, interconnection network and granularity of processors.

#### **1.3.1 BASED ON CONTROL MECHANISM**

Parallel computers are classified as single instruction stream, multiple data stream (SIMD) and multiple instruction stream, multiple data stream (MIMD). Processing units in parallel computers either operate under the centralized control of a single control unit

or work independently. In architectures referred to as SIMD, a single control unit dispatches instructions to each processing unit. Here, the same instruction is executed synchronously by all processing units. Examples of SIMD parallel computers are MasPar MP-1, MasPar MP-2, MPP, DAP and CM-2. Computers in which each processor is capable of executing a different program independent of the other processors are called multiple instruction stream, multiple data stream (MIMD) computers. Example of MIMD computers are Cosmic Cube, nCUBE-2, iPSC, CM-5, Paragon XP/S and Cray-T3D.

SIMD computers require less hardware than MIMD as they have only one global control unit. SIMD also requires less memory as only one copy of the program needs to be stored. In contrast, MIMD computers store the program and operating system at each processor. SIMD computers are naturally suited for data parallel programming. Though individual processor in an MIMD computer are more complex, general purpose microprocessors may be used. Hence, due to the economy of scale, processors in MIMD computers are both cheaper and more powerful than processors in SIMD computers.

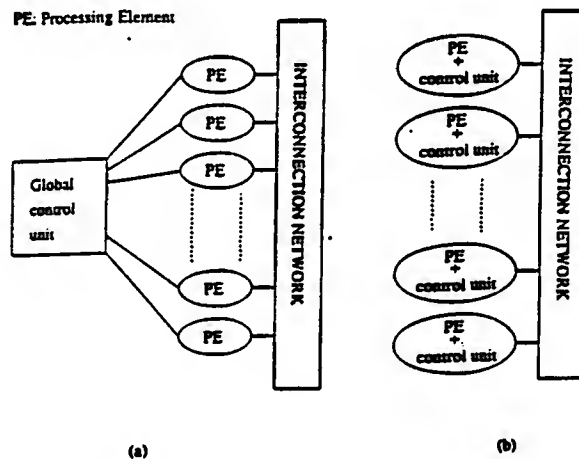
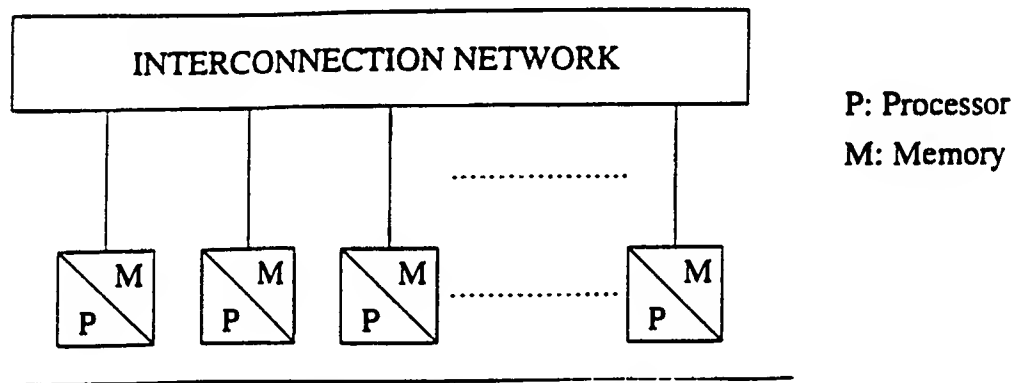


Figure 1-1: (a) Typical SIMD architecture and (b) Typical MIMD architecture

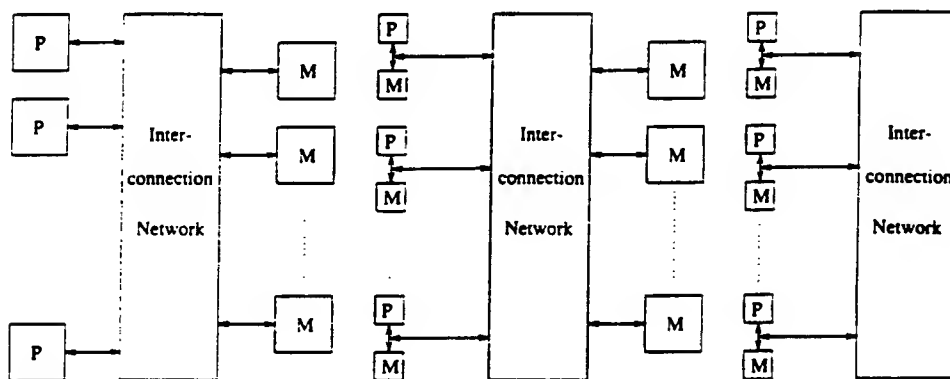
### 1.3.2 BASED ON ADDRESS-SPACE ORGANIZATION

Solving a problem on an ensemble of processors requires interaction among processors. Message Passing and shared address space architectures provide two different means of processor interaction. In a message passing architecture, processors are

connected using a message passing interconnection network. Each processor has its own memory called local memory, which is accessible only to that processor. Processors can interact only by passing messages. It is also referred to as distributed memory architecture. Examples include nCUBE-2, CM-5, Cosmic Cube and Paragon XP/S. The shared address space architecture provides hardware support for read and write access by all processors to a shared addressed space. Most shared address space computers contain a shared memory that is equally accessible to all processors through an interconnection network. These architectures are called shared memory parallel computers. Examples include C.mmp and NYU Ultracomputer. A drawback of these architectures is that the bandwidth of the interconnection network must be substantial to ensure good performance. The Cray-T3D is a logically shared, physically distributed memory.



**Figure 1-2: Message Passing architecture**



**Figure 1-3: Shared-address space architectures**

### **1.3.3 BASED ON INTERCONNECTION NETWORKS**

Shared address space computers and message passing computers can be constructed by connecting processors and memory units using a variety of interconnection networks. Classification of parallel computers based on interconnection networks are static and dynamic. Static networks consists of point to point communication links among processors and are also referred to as direct networks. Static networks are typically used to construct message passing computers. Dynamic networks are built using switches and communication links. Communication links are connected to one another dynamically using switching elements to establish paths among processors and memory banks. Dynamic networks are referred to as indirect networks and are normally used to construct address space computers.

### **1.3.4 BASED ON INTERCONNECTION NETWORKS**

A parallel computer may be composed of a small number of very powerful processors or a large number of relatively less number of processors. Processors belonging to the former class are called coarse-grain computers, and those belonging to the later are called fine-grained computers. Examples of coarse-grain computers are Cray-YMP, Cray-C90 which offer a small number of processors each capable of several Gflops and in contrast, a fine grain computer, examples like CM-2, MasPar MP-1 and MasPar MP-2, offer a large number of relatively slow processors. MasPar MP-1 contains up to 16,384 four-bit processors. There is also a class of parallel computers between the extremes and are called medium grain computers, include CM-5, nCUBE-2 and Paragon XP/S.

The granularity of a parallel computer can be defined as the ratio of the time required for a basic communication operation to the time required for a basic computation. Parallel computers for which this ratio is small are suitable for algorithms requiring frequent communication: that is, algorithm in which the grain size of the computation is small. Since such algorithms contain fine-grain parallelism, these parallel computers are often called fine-grained parallel computers. In contrast, parallel computers for which this ratio is large are suited for algorithms that do not require frequent communication. These

computers are referred to as coarse-grain computers. According to this criterion, multi computers such as the nCUBE 2 and Paragon XP/S are coarse grain computers, whereas multiprocessors such as the C.mmp, TC-2000 and KSR-1 are fine grain parallel computers. The Cray-T3D is a moderately coarse parallel computer.

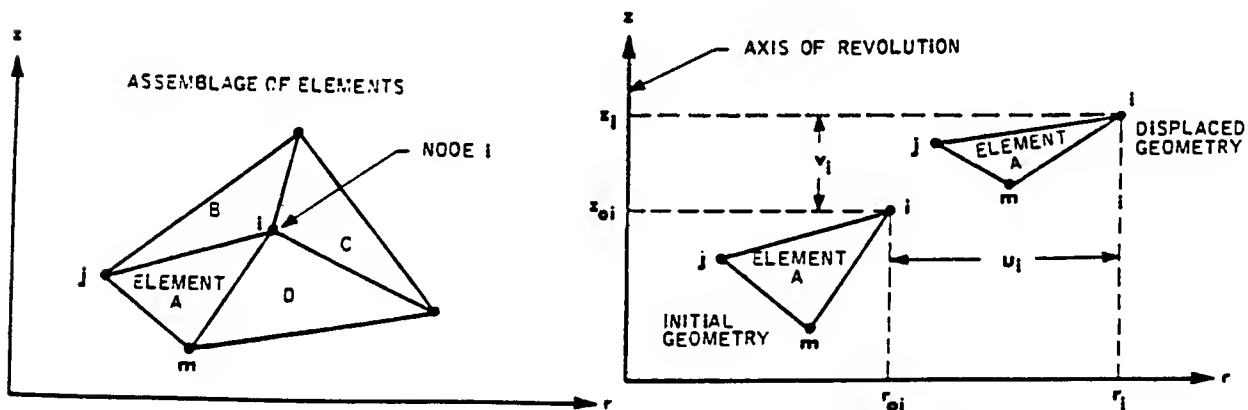
## **2. EPIC RESEARCH HYDRO CODE**

### **2.1 EPIC THEORY**

There are many different computer codes available to calculate dynamic response to impact. For impact problems involving elastic-plastic flow with large displacements, the solutions have most often been obtained with Lagrangian code. The EPIC code is also an Lagrangian FEA code developed by Dr. G.R. Johnson et al, of Alliant Techsystems Inc. It takes advantage of the fact that triangular or tetrahedral element formulation is better suited to represent the severe distortions than is the traditional quadrilateral or hexahedral finite difference methods [4,5].

The finite element method is implemented in the following steps:

- The geometry of the problem is represented with elements of triangular cross-section having specific material characteristics.



**Figure 2-1: Geometric properties of triangular elements**

- The distributed mass at the nodes is lumped. The initial velocities to represent the motion at impact is assigned.

The lumped mass in each of three node is:

$$M_i = M_j = M_m = (1/3)V_o\rho_o$$

where  $V_o$  and  $\rho_o$  are initial volume and density of the element.

- Numerical integration loop works as follows:
- ◊ The strain and strain rates in the elements are determined.

$$\varepsilon = \left[ (2/9) \left\{ (\varepsilon_r - \varepsilon_z)^2 + (\varepsilon_r - \varepsilon_\theta)^2 + (\varepsilon_z - \varepsilon_\theta)^2 + (3/2)\gamma_{rz}^2 \right\} \right]^{1/2}$$

where  $\varepsilon$  is equivalent strain,  $\varepsilon_r, \varepsilon_\theta, \varepsilon_z$ , and  $\gamma_{rz}$  are strains relative to system axes.

- ◊ Stresses in the element are determined. The stresses consists of elastic stresses, plastic deviator stresses, hydrostatic pressure, and artificial viscosity.

Elastic stresses are obtained by Hooke's law.

$$\sigma_r = \lambda \varepsilon_v + 2G \varepsilon_r - Q$$

$$\sigma_z = \lambda \varepsilon_v + 2G \varepsilon_z - Q$$

$$\sigma_\theta = \lambda \varepsilon_v + 2G \varepsilon_\theta - Q$$

$$\tau_{rz} = G \gamma_{rz}$$

where  $\sigma_r, \sigma_z, \sigma_\theta$ , and  $\tau_{rz}$  are radial, axial, circumferential, and shear stresses. and  $\lambda$  and  $G$  are Lamé's elastic constants.

These stresses are combined to form equivalent stress  $\sigma$ .

$$\sigma = \left[ (1/2) \left\{ (\sigma_r - \sigma_z)^2 + (\sigma_r - \sigma_\theta)^2 + (\sigma_z - \sigma_\theta)^2 + 6 \tau_{rz}^2 \right\} \right]^{1/2}$$

This stress represents the overall state of stress within the element.

Plastic flow begins when the elastic strength of materials is exceeded. When this occurs, the normal stresses are obtained by combining hydrostatic pressure with the plastic deviator stresses and the artificial viscosity.

$$\sigma_r = s_r - (P + Q)$$

$$\sigma_z = s_z - (P + Q)$$

$$\sigma_\theta = s_\theta - (P + Q)$$

where  $s_r, s_z,$  and  $s_\theta$  are the plastic deviator stresses,  $P$  is hydrostatic pressure and  $Q$  is the artificial viscosity. The plastic deviator stresses represent the shear strength characteristics of the material.

The hydrostatic pressure is determined from the Mie-Gruneisen equation of state.

$$P = (K_1 \mu + K_2 \mu^2 + K_3 \mu^3) [1 - \Gamma \mu / 2] + \Gamma \rho_o E$$

where

$$\mu = (\rho / \rho_o) - 1 = (V_o / V) - 1$$

The specific internal energy,  $E$ , is obtained from the work done on the element by various stresses,  $K_1, K_2,$  and  $K_3$  are material dependent constants and  $\Gamma$  is the Gruneisen coefficient.



The artificial viscosity is combined with the normal stresses to damp out localized oscillations of the concentrated masses. It is applied only when volumetric strain rate is negative.

$$Q = C_L[(\lambda + 2G)\rho A]l/2|\dot{\epsilon}_v| + C_0^2 \rho A(\epsilon_v)^2 \text{ for } \dot{\epsilon}_v < 0$$

$$Q = 0 \text{ for } \dot{\epsilon}_v \geq 0$$

where recommended value of coefficients are  $C_L = 0.5$  and  $C_0^2 = 4.0$ .

- ◇ Equivalent concentrated forces that act on nodal masses is determined. The radial and axial forces,  $F_{ri}$  and  $F_{zi}$ , acting on node  $i$  of an element are:

$$F_{ri} = -\pi r[(z_j - z_m)\sigma_r + (r_m - r_j)\tau_{rz}] - (2/3)\pi A \sigma_\theta$$

$$F_{zi} = -\pi r[(r_m - r_j)\sigma_z + (z_j - z_m)\tau_{rz}]$$

where the nodal coordinates represent the displaced geometry.

- ◇ Integration time increment is determined.  
Maintaining a numerically stable solution for dynamics problem is generally accomplished by using a numerical integration time increment which is sufficiently less than the lowest period of vibration of the system.
- ◇ Equation of motion to the nodes are applied for integration time increment. The equations of motion can be numerically integrated by assuming a constant acceleration for each time increment. The radial acceleration for each time increment. The radial acceleration of node  $i$  at time ' $t$ ' is:

$$\ddot{u}_i = (\sum F_{ri} / \sum M_i)$$

The new displacement at time  $t + \Delta t$  is

$$u_{t+\Delta t} = u_t + \dot{u}_t \Delta t + (1/2)\ddot{u}_t (\Delta t)^2$$

and the new velocity is

$$\dot{u}_{t+\Delta t} = \dot{u}_t + \ddot{u}_t \Delta t$$

The equation of motion for the axial direction have a similar form. After the equation of motion are numerically integrated, an integration cycle is complete.

- The numerical integration loop is repeated until the time of interest is elapsed.

### **3. PARALLEL PROGRAMMING TECHNIQUES ON CRAY-T3D**

#### **3.1 BACKGROUND**

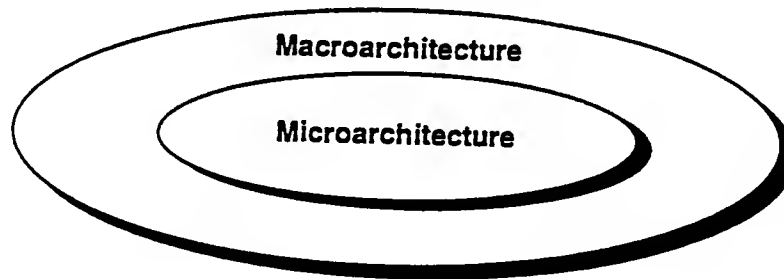
The first step in parallel programming is the development of the parallel algorithm to solve a problem. There are two different approaches to algorithm development. The first approach is when a sequential algorithm is already present and with minor modifications can be changed into a parallel program using parallel programming paradigms. The first approach may not succeed in many cases where the sequential algorithm has too many bottlenecks. In that case, parallel algorithm has to be developed which may have totally different approach as the sequential algorithm.

To run a parallel algorithm on a parallel computer, one needs to implement them on a programming language. In addition to providing all the functionality of a sequential language, a language for programming parallel computers must provide mechanisms for sharing information among processors. It must do so in a way that is clear, concise, and is readily accessible to the programmer. Different parallel programming languages enforce different programming paradigms. The variations among paradigms are motivated by several factors. First, there is a difference in the amount of effort invested in writing parallel programs. Some languages require more work from the programmer, while others require less work but yield less efficient code. Second, one programming paradigm may be more efficient than others for programming on certain parallel programming architectures. Third, various applications have different types of parallelism, so different programming languages have been developed to exploit them.

#### **3.2 CRAY-T3D ARCHITECTURE OVERVIEW**

The introduction of Cray-T3D by CRI in late 1993 has been a significant event in the field of massively parallel supercomputing. The T3D promises a major advance in highly parallel hardware with respect to low latency (1 micro second) and high bandwidth (125 MB/sec) interconnect. The Cray-T3D is scalable to 2048 processor elements and 300 Gflops. It is a multiple instruction multiple data (MIMD) architecture machine. It has a

logically shared and physically distributed DRAM. Cray-T3D has a two-tier architecture consisting of macro architecture and micro architecture.



**Figure 3-1: Cray-T3D two tier architecture**

### **3.2.1 MICRO ARCHITECTURE**

The micro architecture will vary as technologies advance to achieve Tflops of sustained performance. Micro architecture refers to the microprocessor chip used in the Cray-T3D. For its first generation machines (T3D), CRI choose DEC chip 21064 (Alpha chip) for its performance, features, technology and availability. The alpha chip consists of four main components IBOX, EBOX, FBOX and ABOX.

**(a) Central control unit (IBOX):** The IBOX performs instruction fetch, resource checks, and dual instruction issue to the EBOX, ABOX and FBOX or branch unit. It handles pipeline stalls, aborts and restarts.

**(b) Integer execution unit (EBOX):** The EBOX contains a 64-bit fully pipelined integer execution data path including adders, logic box, barrel shifter, byte extract and mask, and independent integer multiplier. In addition, it contains a 32 entry 64-bit integer register file.

**(c) Floating point unit (FBOX):** The FBOX contains a fully pipelined floating point unit and independent divider, supporting both IEEE and VAX floating point data types.

(d) **Load/Store or address unit (ABOX):** The ABOX contains five major sections: address translation data path, load silo, write buffer, data cache (DCACHE) interface and external bus interface unit.

The alpha chip uses a seven stage pipeline for integer operation and memory reference instructions, and a six stage pipeline floating point operations instructions. The IBOX maintains all pipelines stages to track outstanding register writes.

It also contains two on-chip caches: data cache (DCACHE) and instruction cache (ICACHE). The chip also supports secondary cache, but it is not used in the version utilized in the T3D. The data cache contains 8 KB and is a write through, direct mapped, read-allocate physical cache with 32-byte blocks. The data cache is "direct mapped". A direct mapped cache has only one image of a given cache line. It is "read allocate" which means that entries into the cache only happen as a result of a cacheable load from local memory. During a cache hit data is loaded into register from DCACHE and during a cache miss one cache line is loaded from DRAM.

The instruction cache is 8 KB and is a physical direct-mapped cache with 32-byte blocks. The Alpha chip supports secondary cache built from off the shelf static RAMs although it is not used in the T3D. The chip directly controls the RAM s using its programmable secondary cache interface, allowing each implementation to make its own secondary cache speed and configuration tradeoffs.

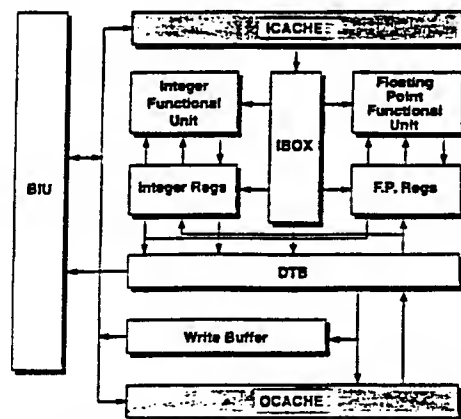


Figure 3-2: Chip block diagram

### **3.2.2. SINGLE PE OPTIMIZATION**

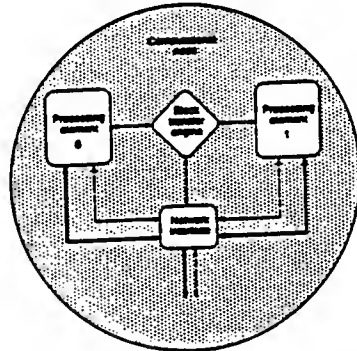
Optimization of a code for a single PE on the Cray-T3D is more difficult than for the other Cray PVP system processor like C90 because of the following: optimizations are state dependent, data locality is always the issue, bandwidth is a limitation factor, not as many functional units are pipelined, compilers and various tools are not as matured as those available on C90. The following are the problems associated with the Alpha chip used in Cray-T3D: (a) all memory operations stall upon cache miss, (b) the slow external bus makes the DRAM bandwidth sub-optimal, (c) there are no integer to floating point or SQRRT instructions, (d) divide and integer multiply are not pipelined. A division operation produces one result every 64 clock periods and integer multiply produces one result every 20 clock periods.

Every DRAM request results in a cache line load of four 64-bit words - one for the actual request and the other three words which are mapped to the same cache line. Aligning data on the same cache line boundary (word 0 of any cache line) enhances the performance. The cache alignment can be done by using a compiler directive `C DIR$ CACHE_ALIGN`. Performance can also be enhanced by scalar replacement, by holding the value of a temporary scalar in a register to reduce the number of memory accesses. Cache utilization can also be enhanced by loop interchange so that stride in the inner loop is one. Large stride in the inner loop causes cache misses. The DRAM memory of the alpha chip is interleaved and one should ensure page boundary alignment. Page hit occurs when either current or previous references are to the same even or same odd page, or current and previous references have different chip select (cs) bit. Page miss occurs when current and previous references are to the different odd pages. Page hits take 8 clock periods whereas a page miss takes 22 clock periods.

### **3.2.3. MACRO ARCHITECTURE**

The macro architecture of the Cray-T3D is based on 3D-torus. It will remain the same in all the three phases of the MPP project. The macro architecture will be stable from one

generation to the next in order to preserve the applications development investment of the users. The 3D-torus is a three dimensional grid with periodic boundary conditions. The 3D-torus was chosen for various reasons: scaling properties, low latency, high bisection bandwidth and low contention. Each node in this topology consists of two PEs, Block transfer engine, and support circuitry which includes Data Translation Buffers (DTB), Message Queue Control (MQC), Data Prefetch Queue (DPQ), Atomic Swap Control (ASC), Barrier Synchronization Registers (BSR), and PE control.



**Figure 3-3: Node architecture**

Each computational mode has two identical PEs which function independently of each other. Each node has support circuitry including but not limited to network interface, network router and block transfer engine. The network interface formats the information and the network router deformats it before sending it to PE0 or PE1. Block transfer engine (BTE) is asynchronous and is shared by two PEs. It can move data independently without involving either the local PE or the remote PE. It also provides gather scatter functionality in addition to data pre-fetch with a constant stride. It can transfer up to 64 K words and can be used to select PE number and memory offset bits using the virtual global memory address facility. The use of BTE requires making system calls. It also involves performing local work first and double buffering the remote data transfers and working on those buffers, but however, the start up time for BTE is very high.

The 3D-torus network, which is a high interconnection network, connecting the nodes operates at 150 Mhz, identical to the clock of the alpha chip used in the node. This leads to low latencies for communication between nodes. It uses "dimensional order routing"

for propagation of messages. The network channels are 16 bits wide and can send simultaneously bi-directional in all three directions (X, Y and Z). The bandwidth of the channel is 300 Mbytes/s. The bisection bandwidth with 1024 PEs is 75 GBytes/s. For node to node message passing, the minimum measured latency is 1.3 microseconds. The network has hardware synchronization primitives which lends to fast synchronization or barriers. The T3D network transmits system control information and user data. The control packets vary in size from 6 to 16 bytes. The data packets range in size from 16 bytes to 52 bytes. The amount of data in these packets is 8 or 32 bytes with the remainder being header and checksum information. The headers and checksums contribute to a load factor which affects attainable data transfer rates [6].

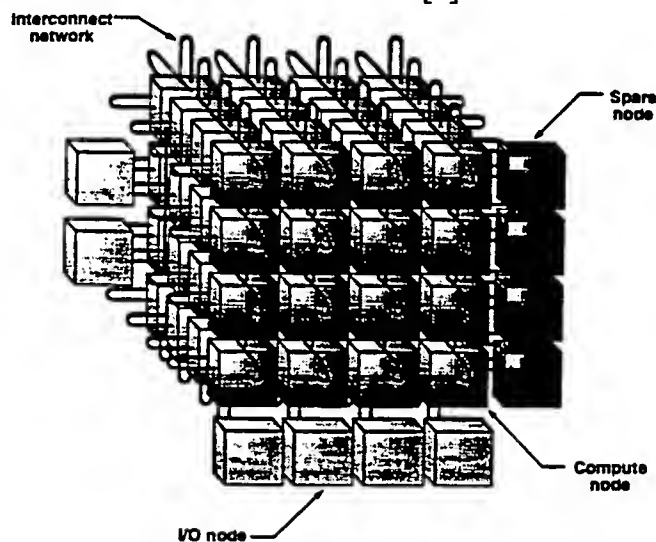


Figure 3-4: Cray-T3D topology

### **3.3 MPP PROGRAMMING MODEL**

The MPP programming model [7] for the Cray-T3D system supports several styles of programming - data parallel, global address, work sharing and message passing. These styles may be individually used or combined in the same program. This model allows the user a range of control over the MPP hardware. This range extends from a low level control in which the programmer makes almost all of the decisions about how data and work are partitioned and distributed, to a high level of control where the programmer identifies where parallelism is located and lets the system determine best how to exploit it. The important elements of this programming model are access and placement of data,



parallel execution, local execution, work sharing, synchronization primitives, sequential I/O, subroutine interfaces and special intrinsic functions.

### **3.3.1. DATA PARALLEL MODEL**

The MPP programming model distinguishes data objects into two categories :

- (1) private data (PE\_PRIVATE), that are private to a task
- (2) shared data (SHARED), that are shared among all the tasks.

Private data objects reside on each PE, rather than spreading one copy over all of them. They are not accessible to any other task. The task that references a private object references its own private version of that object and therefore it is possible for private data objects associated with different PEs to have different values.

Shared data objects, on the other hand, are accessible to all tasks. They are not replicated and in case of arrays be distributed across multiple Pes.

In data parallel programming, data such as scalar or array are distributed over the memories of the PEs working on the program. In this programming model, the goal is to let as many PEs as possible perform on its own data (residing in its memory) rather than working on the data that is residing in another PE's memory.

In CRAFT, the data distribution is indicated by the compiler directives PE\_PRIVATE and SHARED. Data that are not explicitly declared to be shared is, by default, private data. Variables and arrays can be explicitly declared as private with the PE\_PRIVATE directive.

```
CDIR$ PE_PRIVATE var1, var2 ... varn
```

All private data objects may be DATA initialized except those that occur in blank common, dummy arguments, automatic arrays, and those whose size is a function of N\$PES (number of tasks).

The shared data objects are declared with the SHARED directive.

```
CDIR$ SHARED array1(dist1),array2(dist2) ...arrayn(distn)
```

The shared directive names the variables that are to be shared data objects and specifies the distribution across the PEs. Shared data object's distributions fall into two categories : shared scalars and dimensional distribution. Scalar variables are always allocated on a single PE, which may differ for different processor elements. Dimensional distribution includes the following: Cyclic distribution, Generalized distribution, Block distribution and Degenerate distribution.

Cyclic distribution (:BLOCK(1)) assigns one element of shared array to each PE, returning to the first PE when every PE has an element. Generalized distribution (:BLOCK(n)) assigns blocks of 'n' elements of the array to successive PEs, where 'n' has to be an integer power of 2. The block distribution (:BLOCK) divides an array dimension into N\$PES blocks and allocates one block to each PE. The block size equals to array size divided by N\$PES. Degenerate distribution (:) forces an entire dimension to be allocated on a single PE.

### **3.3.2. WORK SHARING**

Executing the statements of program in parallel, and in the same PEs in which the data is distributed achieves higher performance for the Cray MPP system. Work sharing is achieved primarily by two ways: automatic arrays and shared DO loops.

Fortran array syntax or automatic arrays is one way to distribute work. A Fortran statement using array syntax and involving shared arrays encountered in a parallel region

causes all processors to execute the statement. The compiler maximizes data locality i.e., the work is such distributed that tasks execute on its local data.

```
DIMENSION Z1(64), Z2(64), VNEW(64)
CDIR$ SHARED Z1(:BLOCK), Z2(:BLOCK), VNEW(:BLOCK)
.....
VNEW(I) = Z1(I) - Z2(I)
```

DOSHARED directive is the second way of achieving work sharing. As loops do not create parallelism, work sharing of DO loops is achieved by distributing iterations across all available tasks. Each task is assigned a set of iterations of a shared loop to execute. Shared loops do not guarantee the order in which iterations will be executed and lets the system execute iterations concurrently. There is an implicit barrier synchronization at the end of a shared loop. The example for DOSHARED directive from subroutine VOLUME is as follows:

```
CDIR$ DO SHARED (I) ON VNEW(I)
DO 10, I = 1, LNL1
VNEW(I) = Z1(I) - Z2(I)
10 CONTINUE
```

Private loops can be inside and outside the shared loop, but the shared loop must be tightly nested, the inner shared loop is executed as a private loop. The distribution mechanism for a shared loop affects program performance rather than correctness. Proper choice of iteration alignment provides higher degree of locality (when references in the iteration are close together). The aligned distribution mechanism is designed to place iterations within tasks on PEs where the references reside.

A private loop is executed only by the task that invokes it and no work is shared between tasks. Private loops define program behavior by defining the behavior of the

individual tasks. Private loops have exactly the same semantics as loops in standard Fortran. No special syntax is required to specify a loop as private, as it is the default.

### **3.3.3. SHARED TO PRIVATE COERCION**

The cardinal rule for distributed memory machines is to exploit data locality i.e, work with local data and avoid communication as much as possible. Performance without communication far exceeds that with communication.

There are two paradigms of CRAFT with no interprocessor communication. The highest performance is attained by shared to private coercion and the next paradigm is the PE\_RESIDENT directive.

In shared to private coercion, an actual argument declared as shared array is passed to a corresponding dummy argument declared as a private array. This causes each PE to pass only its own data to the subroutine, which leads to the subroutine accessing array elements that are strictly local to the executing PE as private data without additional overhead.

The example illustrated here is from the STRAIN subroutine:

```
PROGRAM START_STRAIN
INTEGER L1, LN, M, LNL1
REAL Z1DOT(MXLB), Z2DOT(MXLB), HMIN(MXLB), EZDOT(MXLB),
2      EXDOT(MXLB), EXYDOT(MXLB), EYDOT(MXLB)
.....
CDIR$ GEOMETRY GG(:BLOCK(1))
CDIR$ SHARED (GG) :: Z1DOT, Z2DOT, HMIN, EZDOT, EXDOT, EXYDOT,
2      EYDOT
.....
LNL1 = (LN-L1+1)/N$PES
```

```

      CALL STRAIN(L1, LN, EXDOT, EXYDOT, EYDOT, EZDOT,
*           HMIN, Z1DOT, Z2DOT,...LNL1)
      ....
      END
      SUBROUTINE STRAIN(L1, LN, EXDOT, EXYDOT, EYDOT, EZDOT,
*           HMIN, Z1DOT, Z2DOT,...LNL1)
C      STRAIN computes strain rates
      REAL Z1DOT(*), Z2DOT(*), HMIN(*), EXDOT(*), EZDOT(*),
2      EXYDOT(*), EYDOT(*)
      ....
      IF (IGEOM.EQ.1) THEN
          DO 10, I=1,LNL1
              EZDOT(I) = (Z1DOT(I)-Z2DOT(I))/HMIN(I)
              EXDOT(I)=0.0
              ....
10      CONTINUE
      ....
      RETURN
      END

```

In this example, the variables EXDOT, EYDOT, EXYDOT, EZDOT, Z1DOT, Z2DOT and HMIN are defined as shared variables in the calling program, but are defined as private variables in the called subroutine STRAIN. This causes each PE to pass only its own data to the subroutine like the first element of these arrays are located in PE0, the second element of these arrays are located in PE1 and so on. So, when DO loop is executed, all the executable statements work on local data. This, reduces the communication time and improves the performance of the parallel code.

## **4. PERFORMANCE OF INDIVIDUAL SUBROUTINES**

### **4.1 PERFORMANCE MODELING AND SEALABILITY ANALYSIS FOR PARALLEL SYSTEM**

A sequential algorithm is evaluated in terms of its execution time, expressed as a function of the size of its input. The execution time of a parallel algorithm depends not only on input size but also on the architecture of the parallel computer and the number of processors. A parallel system is a combination of an algorithm and the parallel architecture on which it is implemented. The performance of a parallel program takes into account execution time, scalability of computational kernels, the mechanisms with which data is generated, stored, transmitted over networks, moved to and from disks, and passed between different stages of a computation. Metrics used to measure performance include execution time, parallel efficiency, memory requirements, throughput, latency, input/output rates, network throughput, design costs, implementation costs, verification costs, potential for reuse, hardware requirements, hardware costs, maintenance costs, portability and scalability. The relative importance of these diverse metrics will vary to the nature of problem at hand. A specification may provide hard numbers for some metrics, require that others be optimized, and ignore yet others. For example, the design specification for an operational weather forecasting system may specify maximum execution time (like, the forecast must be complete within four hours), hardware costs, and implementation costs, and require that the fidelity of the model be maximized within these constraints. In addition, reliability is of particular high importance, as may be scalability to future generation of computers. For a different application like image processing, one is not concerned with total time required to process a certain number of images but rather with the number of images that can be processed per second (throughput) or the time that it takes a single image to pass through the pipeline (latency). Throughput would be important in a video compression application, while latency would be important if the program formed a part of a sensor system that must react in real time to events detected in a image stream.

For the EPIC hydro code, though the wall clock time is important, the most important metrics one is concerned about is the cost of computation. Comparing cost of computation between the Cray-YMP and Cray-T3D, using 32 processors on the Cray-T3D is cheaper than running an application code on single YMP processor. So, the goal of parallelizing the EPIC code is to run it on a more cost effective parallel system.

## **4.2 PERFORMANCE METRIES IN PARALLEL SYSTEMS**

Some of the metrics that are commonly used to measure the performance of parallel systems are described below:

### **4.2.1 RUN TIME**

The serial run time of the program is the time elapsed between the beginning and the end of its execution on a serial computer. The parallel run time of the time that elapses from the moment that a parallel computation starts to the last processor finishes execution.

### **4.2.2. SPEEDUP**

It is defined as the ratio of the time taken to solve a problem on a single processor to the time required to solve the same problem on a parallel computer with  $p$  identical processors. It is denoted by symbol  $S$ . When evaluating a parallel system, one is interested in knowing how much performance gain is achieved by parallelizing a given application over a sequential implementation. Speedup is measure that captures the relative benefit of solving a problem in parallel.

When a serial computer is used, it is natural to use the sequential algorithm which solves the problem in least amount of time. So, to judge a parallel algorithm fairly, it is compared to the fastest sequential algorithm on a single processor. When the fastest sequential algorithm to solve a problem is not known, or impractical to implement, the fastest known and practical algorithm is chosen as the best sequential algorithm. Then, the performance of the parallel algorithm to solve the problem is compared to the best

sequential algorithm to solve the same problem. So, speedup is defined as the best sequential algorithm for solving a problem to the time taken by the parallel algorithm to solve the problem on p processors. The p processors used by the parallel algorithm are assumed to be identical to the one used by the sequential algorithm.

#### **4.2.3 EFFICIENCY**

An ideal parallel system containing p processors can deliver a speedup equal to p. In practice, ideal behavior is not achieved because while executing a parallel algorithm, the processors cannot devote 100 percent of their time to the computation of the algorithm. Part of the time, required by the processors to solve the problem, is spent in communication. Efficiency is a measure of the fraction of time for which a processor is usefully employed; it is defined as the ratio of speedup to the number of processors. In an ideal parallel system, speedup is equal to p and efficiency equal to one. In practice, speedup is less than p and efficiency is between zero and one, depending on the degree of effectiveness with which the processors are utilized. Denoting efficiency as E,

$$E = S / P$$

#### **4.2.4 COST**

Cost of solving a problem on a parallel system is defined as the product of run time and the number of processors used. Cost reflects the sum of the time that each processor spends solving the problem. The cost of solving a problem on a single processor is the execution time of the fastest known sequential algorithm. A parallel system is said to be cost optimal if the cost of solving a problem on a parallel computer is proportional to the execution time of the fastest-known sequential algorithm on a single processor. Since efficiency is the ratio of sequential cost to parallel cost, a cost optimal parallel system has an efficiency in the order of 1.

#### **4.3 SCALABILITY OF PARALLEL SYSTEMS AND AMDAHL'S LAW**

The number of processors is the upper bound on the speedup that can be achieved by a parallel system. Speedup is one for a single processor, but if more processors are used,



speedup is usually less than the number of processors. This phenomenon is explained by Amdahl's law.

Amdahl's law states that if the sequential component of an algorithm accounts for  $1/s$  of the program's execution time, then the maximum possible speedup that can be achieved on a parallel computer is  $s$ . This is because every algorithm has a sequential component that will eventually limit the speedup that can be achieved on a parallel computer. For example, if the sequential component is 5 percent, then the maximum speedup that can be achieved is 20.

As a consequence of Amdahl's law, the efficiency drops with an increasing number of processors. Secondly, a larger instance of the same problem yields higher speedup and efficiency for the same number of processors, although both speedup and efficiency continue to drop with increasing  $p$ . These two phenomena are common to a large class of parallel systems.

Given that increasing the number of processors reduces efficiency and that increasing the size of the computation increases efficiency, it should be possible to keep the efficiency fixed by increasing both the size of the problem and the number of processors simultaneously. For example, the efficiency of an algorithm of adding 64 numbers using four processors is 0.8. If the number of processors is increased to 8 and the size of the problem is scaled up to add 192 numbers, the efficiency remains 0.8. This ability to maintain efficiency at a fixed value by simultaneously increasing the number of processors and the size of the problem is called scalability of parallel system. The scalability of a parallel system is a measure of its capacity to increase speedup in proportion to the number of processors. It reflects a parallel system's ability to utilize increasing processing resources effectively. The scalability and cost-optimality of parallel systems are related. A scalable parallel system can be made cost-optimal if the number of processors and the size of the computation are chosen appropriately. A good scalable system is one whose efficiency doesn't decrease with increase in the problem size.

#### **4.4 ISOEFFICIENCY OF PARALLEL SYSTEMS**

It is useful to determine the rate at which the problem size must increase with respect to the number of processors to keep the efficiency fixed. For, different parallel systems the problem size must increase at different rates in order to maintain fixed efficiency as the number of processors is increased. This rate determines the degree of scalability of the parallel system.

For scalable parallel systems, efficiency can be maintained at a fixed value (between 0 to 1)

$$E = 1 / (1 + T(W,p) / W)$$

where E is efficiency, T is overhead function, W is problem size and p is the number of processors.

Then, the equation becomes,

$$W = KT(W,p)$$

where  $K = E/(1-E)$  is a constant.

From the above equation, the problem size W can be obtained as a function of p. This function dictates the growth rate of W required to keep the efficiency fixed as p increases. This function is known as the isoefficiency function of the parallel system. The isoefficiency function determines the ease with which a parallel system can maintain a constant efficiency and hence achieve speedups increasing in proportion to the number of processors. A small isoefficiency function means that small increments in problem size are sufficient for the efficient utilization of an increasing number of processors indicating that the parallel system is highly scalable. A large isoefficiency function indicates a poor scalable parallel system.

In a single expression, the isoefficiency function captures the characteristics of a parallel algorithm as well as parallel architecture on which it is implemented. After performing the isoefficiency analysis, one can test the performance of a parallel program on a few processors and then predict its performance on a larger number of processors. However, the utility of isoefficiency function is not limiting to predicting the impact on performance of an increasing number of processors.

#### **4.5 SIGNIFICANCE OF GRAPHICAL PLOTS**

Programmers use parallelism to make their programs solve a single problem in less time or solve larger problems in a fixed time. Substantial amount of effort is spent in presenting the performance in the best possible way. So, it is very important to understand the significance of each graphical plot in terms of performance of the parallel code. In the present work, the following graphical plots were highlighted in the performance of individual subroutines.

- **Number of Processors v/s Wall Clock Time:** In this plot, the wall clock time of the code using multiple processors on T3D and that of YMP is compared. If the wall clock time of 32 processors on T3D is less than the YMP wall clock time, the code or part of the code is cost effective.
- **Number of Processors v/s Speedup:** In this plot, the actual speedup using the particular number of processors is compared with the ideal speedup (no communication assumed!). It gives an idea of the amount of time is spent in communication rather than computation as the processors increase.
- **Number of Processors v/s Efficiency:** In this plot, the efficiency of the algorithm is seen. It gives the idea of the ideal number of processors needed to run the code.

#### **4.6 PARALLELIZATION OF INDIVIDUAL SUBROUTINES**

All the subroutines discussed in the thesis were handled in a particular manner as discussed in paralleling VOLUME. Data parallelism, Work sharing and Shared to Private coercion were the techniques used for parallelizing individual subroutines.

#### **4.6.1. VOLUME**

The VOLUME subroutine computes volumes, volumetric strains and strain rates. The VOLUME subroutine is called by the subroutine ELOOP. To simulate problem for the subroutine the variables and arrays are data initialized in the calling routine, in this case subroutine ELOOP. Among such variables are L1, LN,...X1(I)...etc. As the subroutine in the EPIC program have common include files, the array size is made powers of 2 for compiling on the Cray-T3D.

Firstly, the data in the subroutine is shared using different distribution schemes like (:BLOCK(1)) and (:BLOCK). Work sharing is implemented explicitly by DOSHARED directives in the DO loop. A variable LNL1 is defined which is equal to LN-L1+1 and the original vectorized DO loop index J is eliminated.

The Cray-T3D being a dedicated machine, the real time clock function is used to measure the wall clock time. For, better results, the number of times a subroutine is called is increased, so that the code accumulates some execution time.

Using :BLOCK(1) and :BLOCK data distribution and work sharing directives gave the results tabulated in table 4-1.

**Table 4-1: Comparison of :BLOCK and :BLOCK(1) data distribution directives**

<i>Number of Processors</i>	<i>:BLOCK(1)</i>	<i>:BLOCK</i>
1	2.794	2.543
2	1.121	1.112
4	0.923	0.883

To further improve the performance of the subroutine the shared to private coercion technique was implemented. In this technique, shared arrays in the calling routine are passed to corresponding dummy arguments. Like `x1` is declared in `ELOOP`, but declared private in `VOLUME`. So, in the subroutine `VOLUME` each PE has `x1` in blocks of array-size divided by the number of PEs ( $\text{array-size}/n\text{pes}$  in this case  $\text{LNL1}/n\text{pes}$ ). Due to this DO loops in the subroutine `VOLUME` also indexes upto  $\text{LNL1}/n\text{pes}$ , so that each PE work on its local data and no interprocessor communication takes place. The shared to private coercion vastly improves the performance of the subroutine.

**Table 4-2: Wall clock timing of subroutine VOLUME**

<i>Number of Processors</i>	<i>Wall clock time</i>
YMP	0.763
1	2.298
2	0.900
4	0.513
8	0.281
16	0.183
32	0.136
64	0.118
128	0.108

In the subroutine `VOLUME`, the massively parallel code on 4 PEs runs faster than the vectorized code. This subroutine's MPP implementation is very cost effective. It is also observed that the efficiency of the MPP code is quite high with 32 PEs, which implies good scalable subroutine.

The following figures give an effective representation of parallel performance of the subroutine VOLUME.

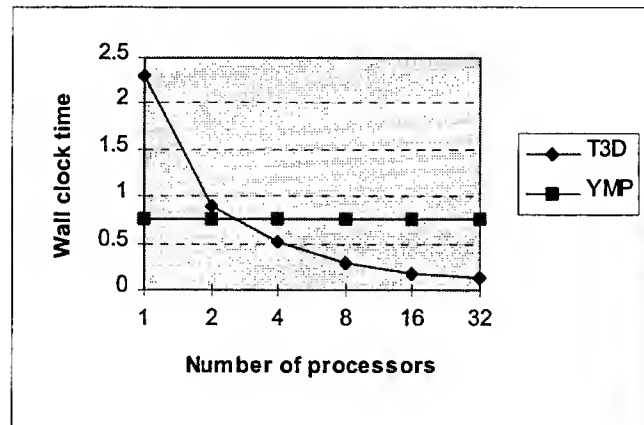


Figure 4-1: VOLUME : Number of processors v/s Wall clock time

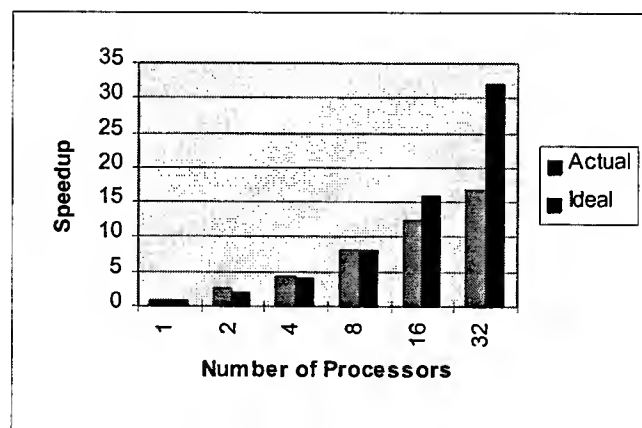
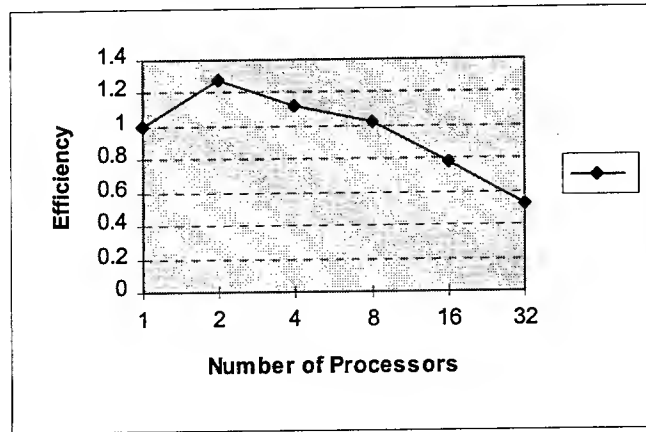


Figure 4-2: VOLUME: Number of Processors v/s Speedup



**Figure 4-3: VOLUME: Number of Processors v/s Efficiency**

#### **4.6.2. EGET**

The EGET subroutine is used to initialize element variables from nodal variables. It is also called by subroutine ELOOP. The variables and arrays are data initialized in the calling subroutine.

To further improve performance, the shared to private coercion was implemented. As in the CRAFT program in the DOSHARED loop contains  $J = L1 + I - 1$ , shared to private coercion cannot be implemented as it is because it contains  $I$  on the right hand side of the assignment statement. So, a variable  $II(I) = I$  was defined in the calling routine and shared to private coercion was implemented.

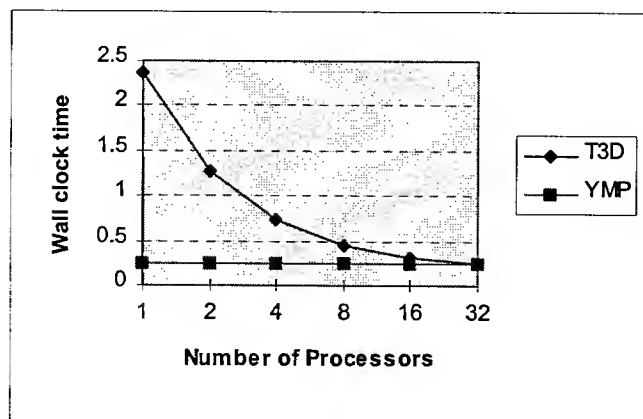
The performance gain by implementing the shared to private coercion principles can be seen in the following table. Not only their is better performance compared to the corresponding CRAFT MPP code, but also is it more scalable.

In the EGET subroutine, the MPP code gives better performance than the YMP code when run on 32 PEs. So, this subroutine is also cost effective, but not as much as subroutine VOLUME. As the number of control statements in the subroutine EGET is

more than subroutine VOLUME, the efficiency of parallel EGET is less. Also, its efficiency is not as high as VOLUME when run on 32 PEs.

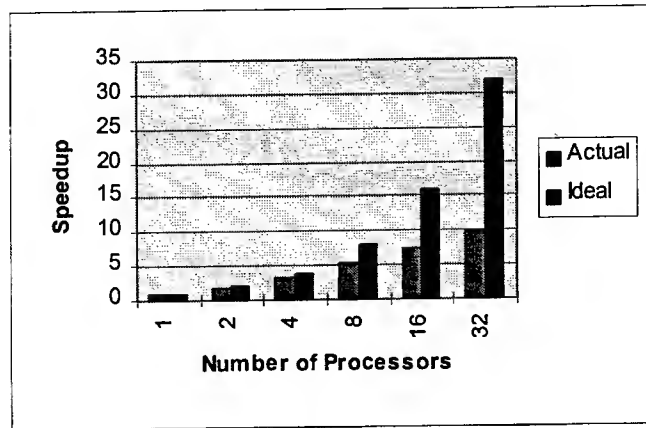
**Table 4-3: EGET: Number of Processors v/s Wall clock time**

<i>Number of Processors</i>	<i>Wall clock time</i>
YMP	0.257
1	2.368
2	1.276
4	0.735
8	0.454
16	0.314
32	0.244

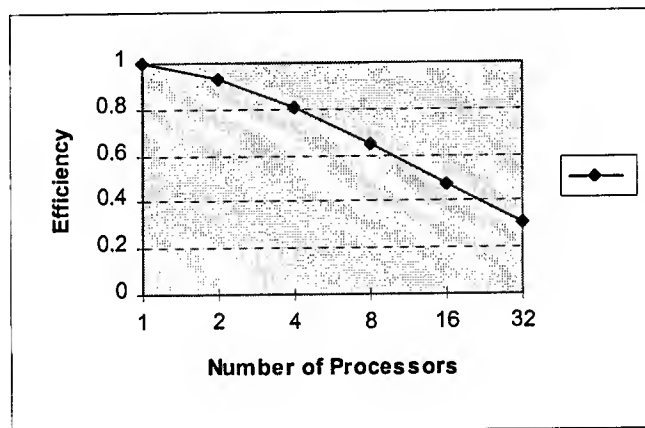


**Figure 4-4: EGET: Number of Processors v/s Wall clock time**





**Figure 4-5: EGET: Number of Processors v/s Speedup**



**Figure 4-6: EGET: Number of Processors v/s Efficiency**

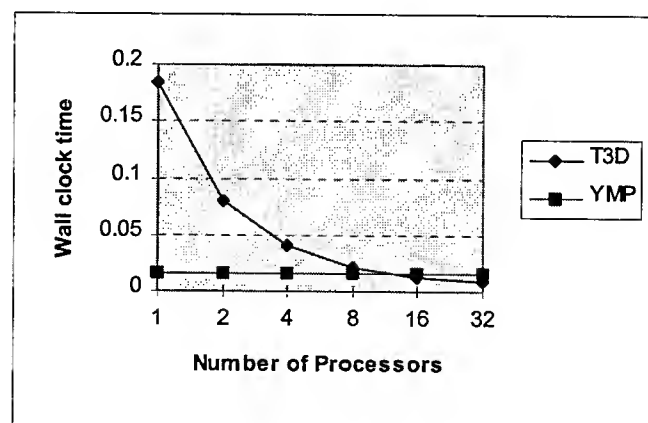
#### **4.6.3. GMCON**

GMCON computes geometric constants for elements in the common core positions. It was parallelized by shared to private coercion technique. The data parallel scheme implemented was :BLOCK. It is seen that the wall clock time of 16 processors in T3D is less than that of YMP, therefore a cost effective routine to parallelize.

The parallel GMCON running on 16 PEs was faster than the YMP vectorized code. The efficiency curve showed that the efficiency of the parallel code was good(>0.6) for 32 PEs and thus could be called a scalable subroutine.

**Table 4-4: GMCON: Number of Processors v/s Wall clock time**

<i>Number of Processors</i>	<i>Wall clock time</i>
YMP	0.0163
1	0.1836
2	0.0788
4	0.4060
8	0.0223
16	0.0133
32	0.0090



**Figure 4-7: GMCON: Number of Processors v/s Wall clock time**

# Supramolecular Multilayer Assemblies with Periodicities in a Submicron Range

Vladimir V. Tsukruk  
Associate Professor

Western Michigan University  
Kalamazoo, MI

Final Report for:  
Summer Research Extension Program  
Wright Laboratory

Sponsored by:  
Air Force Office of Scientific Research  
Bolling Air Force Base, DC

and

Wright Laboratory

February 1997

# **SUPRAMOLECULAR MULTILAYER ASSEMBLIES WITH PERIODICITIES IN A SUBMICRON RANGE**

**(A step toward smart optical filters)**

VLADIMIR V. TSUKRUK

MATERIALS PROGRAM, WESTERN MICHIGAN UNIVERSITY,  
KALAMAZOO, MI 49008

## Abstract.

Organized ultrathin films are the subject of great and growing interest because of their compatibility with future nano-scale technologies. We use the supramolecular engineering approach manipulating a single building unit, a supramolecular assembly, with chemically pre-determined nature of functionality and dimensions to build organized films with large-scale periodicity of multilayer structures. Supramolecular self-assembled films are fabricated by electrostatic layer-by-layer deposition and electrostatic deposition assisted by dip-coating and spin-coating. We use three very different classes of charged polymeric materials: amorphous coiled polyions, dendritic macromolecules, and polymer latex nanoparticles. Self-assembled films are studied by scanning probe microscopy, X-ray reflectivity, ellipsometry, and contact-angle measurements. We demonstrate that replacing unstructured coiled macromolecular chains with organized dendritic supramolecules or "bulk" nanoparticles results in an increase in the growth increment and internal periodicity by an order of magnitude higher than for conventional amorphous films. In-plane ordering can be controlled by deposition time, ionization state, and application of capillary or shearing forces. The routine proposed may be used for formation of supramolecular films with mesoscale periodicity and intriguing optical properties. We present the first truly macroscopically ordered monolayer of charged latex nanoparticles obtained by force assisted electrostatic deposition with lateral sizes extended to a fraction of a millimeter.

# **SUPRAMOLECULAR MULTILAYER ASSEMBLIES WITH PERIODICITIES IN A SUBMICRON RANGE**

## **(A step toward smart optical filters)**

VLADIMIR V. TSUKRUK

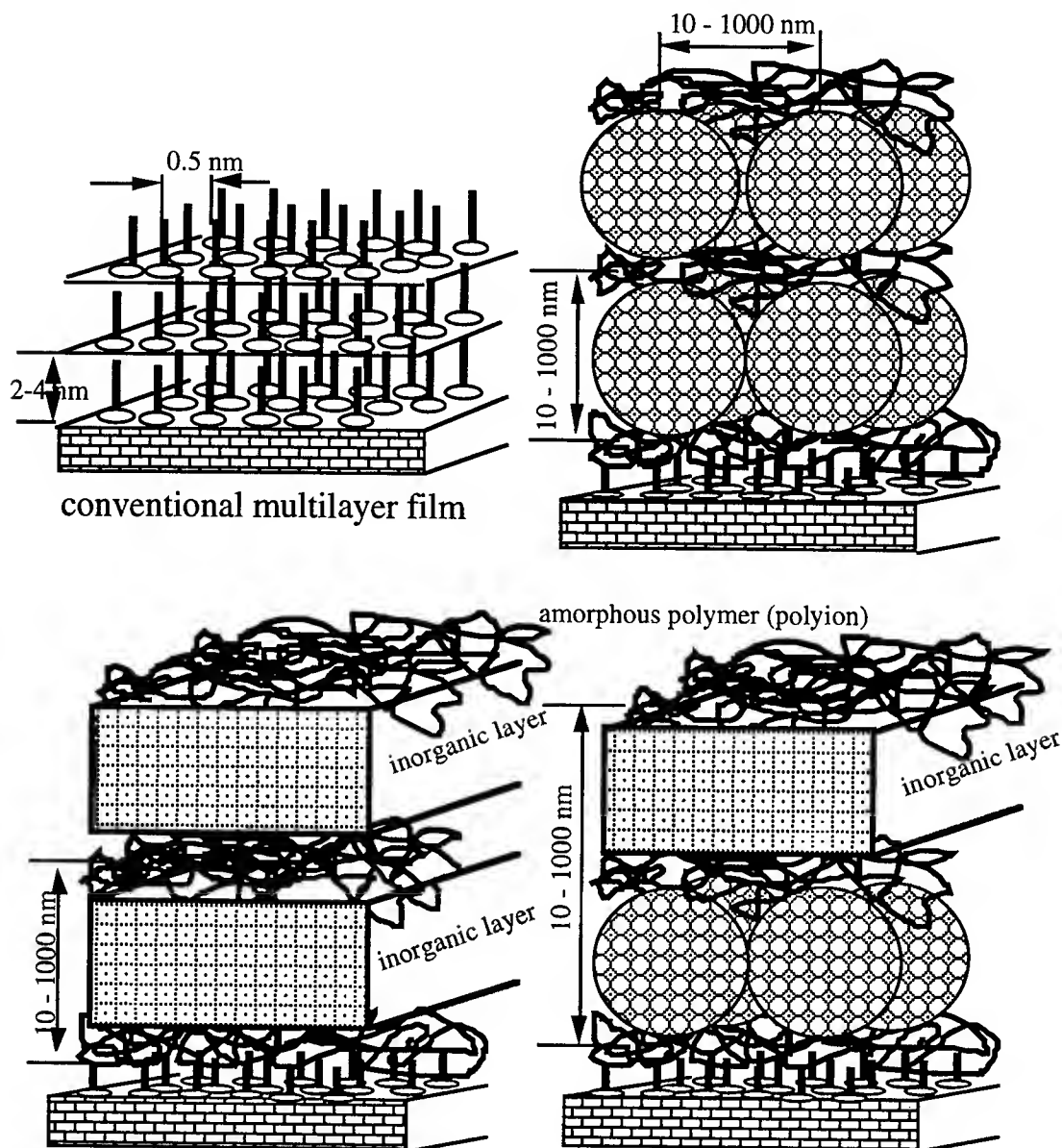
MATERIALS PROGRAM, WESTERN MICHIGAN UNIVERSITY,  
KALAMAZOO, MI 49008

### Introduction.

Organized ultrathin films are the subject of great and growing interest because of their compatibility with future nano-scale technologies. Some examples are integrated optical coatings with a gradient of refractive indices along the normal to the surface plane<sup>1</sup> and microelectromechanical systems modified by boundary lubricants on the molecular scale.<sup>2</sup> Research in these fields focuses on supramolecular functional polymeric materials with suitable physical properties (e. g., non-linear, photochromic, or sensing) and abilities to form organized superstructures at the interfaces.<sup>3-6</sup> The search for new materials with a suitable combination of properties, microstructure, and intermolecular interactions is under the way.

In our studies, we use the supramolecular engineering approach to manipulate with a single building unit, a supramolecular assembly, with chemically pre-determined nature of functionality and dimensions to build organized films with controllable, large-scale periodicity of multilayer structures (see Scheme 1 for general comparison).<sup>5, 6</sup> Modulation of internal structural periodicity of polymer films at a submicron scale, which results in accompanying variation of the refractive index, is an interesting route toward the next generation of optical reflective filters.

Currently, several approaches exist for fabrication of organized molecular assemblies from functional macromolecular materials. Electrostatic layer-by-layer deposition from dilute solutions which exploits Coulombic interactions between oppositely charged molecules has become a widely used method since 1991.<sup>7</sup> It has been shown that this approach can be used to build multilayer films (hundreds of layers) with various combinations of molecular fragments, organic and inorganic layers, latexes, molecules with switchable conformation, biomolecules, photochromic molecules, and conductive polymers.<sup>3-8</sup>



Scheme 1. Possible architectures of supramolecular films.

The layer-by-layer self-assembling process of amorphous polyions in its initial stage requires special attention. The mechanical and temporal stability of the first molecular layers tethered to a solid substrate is a critical element to the homogeneity of thicker films. A gradient of molecular ordering across the first several molecular layers is observed for various molecular films.<sup>7-9</sup> This phenomenon is usually related to healing of substrate inhomogeneities and non-equilibrium behavior. Formation of non-equilibrium surface morphologies and inhomogeneous coverage is caused by the competition of the kinetics of polymer chain adsorption and their surface diffusion. It is speculated that assembly of polyions on charged

surfaces is indeed a two stage process: macromolecular chains are anchored to the surface by some segments during the short initial stage and then relax to a dense packing during the long second stage of self-assembly.

In the present report, we summarize our observations on the formation of the supramolecular self-assembled films by electrostatic deposition of three very different classes of polymeric materials: amorphous coiled polyions, dendritic macromolecules, and polymer latex nanoparticles (Figure 1). We make preliminary conclusions on feasibility of supramolecular engineering approach for building organized polymer films with mesoscale periodicity.

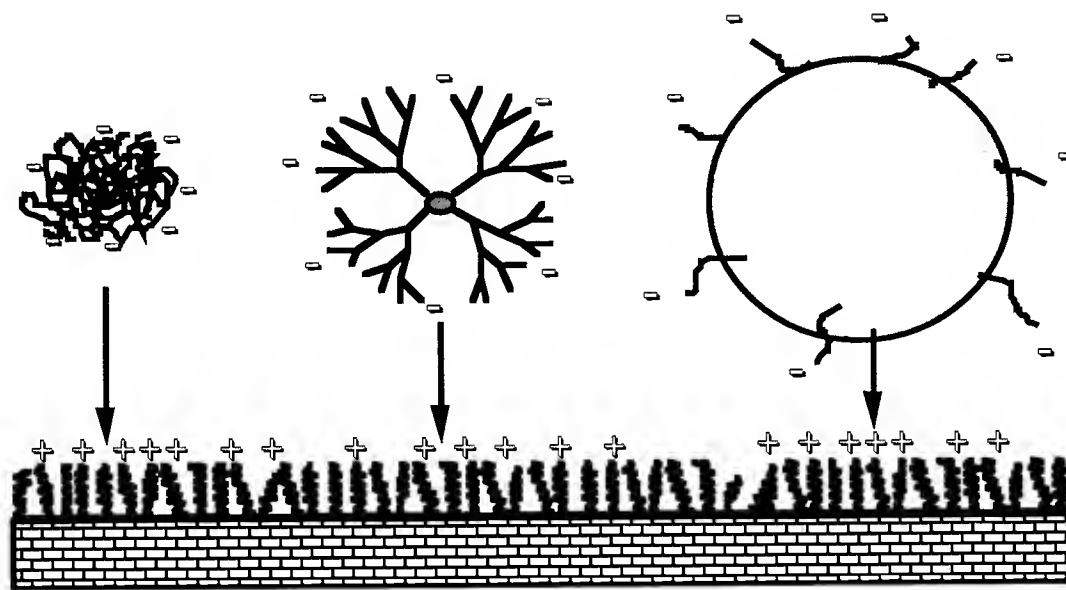


Figure 1. Electrostatic adsorption of charged polymer coils, dendrimers, and nanoparticles on oppositely charged surfaces.

### Experimental

Negatively-positively charged pairs of polystyrene sulfonate (PSS) and polyallylamine (PAA) (Figure 2) <sup>10</sup>, PS latexes of 20 - 200 nm in diameter with amino and carboxy surface groups (Table 1) <sup>11</sup>, and polyamidoamine dendrimers with surface amine groups for even generations and carboxylic groups for odd generations (G3.5 - G10) (Figure 3) <sup>12</sup> were selected for this study.

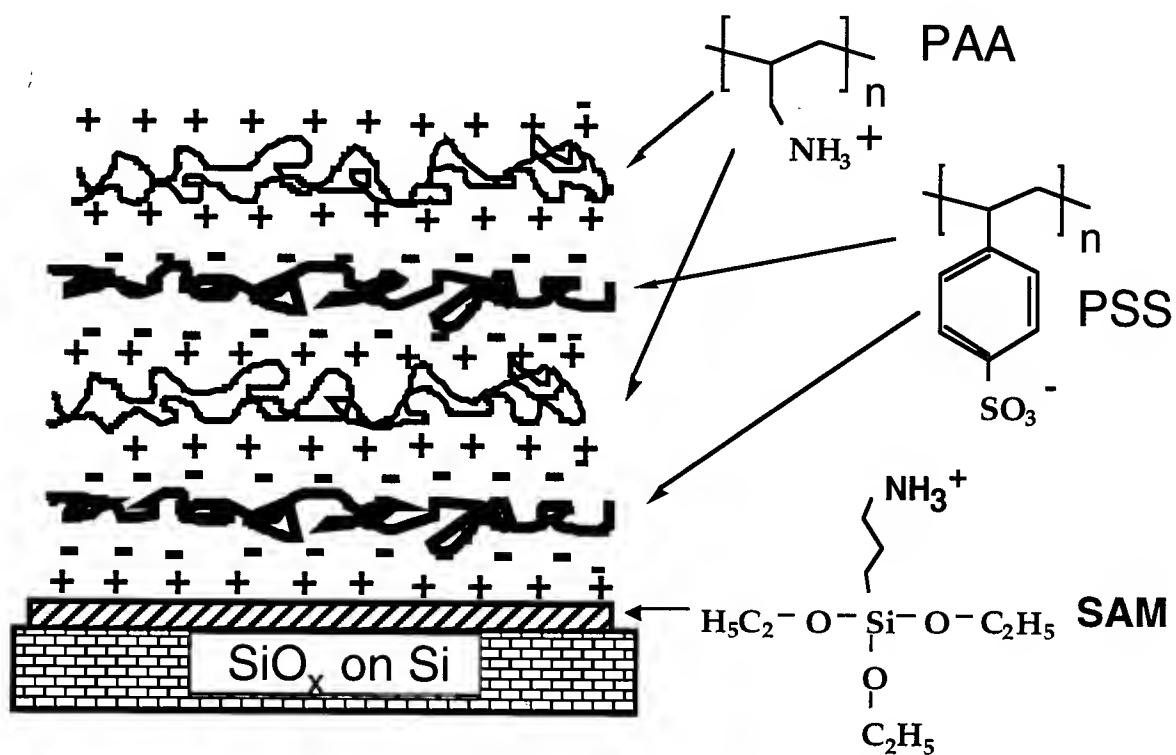


Figure 2. Self-assembled films and PSS/PAA formulas.

Table 1. Latex characteristics and notations

Latex	Mean Particle Diameter (D), nm	Standard Deviation of D, %	Surface Charge Type and Density	Type of the Surface	Notation
Amidine-modified PS	20	23.3	positive $1.7 \mu\text{C}/\text{cm}^2$	hydrophobic	AL20
Carboxyl-modified PS	20	15.3	negative $17.7 \mu\text{C}/\text{cm}^2$	hydrophilic	CML20
Sulfate PS	40	13.9	negative $2.5 \mu\text{C}/\text{cm}^2$	hydrophobic	SL40
Amidine-modified PS	190	1.5	positive $8.3 \mu\text{C}/\text{cm}^2$	hydrophobic	AL190
Carboxyl-modified PS	190	2.7	negative $221 \mu\text{C}/\text{cm}^2$	hydrophilic	CML190
Sulfate PS	200	4.6	positive $1.1 \mu\text{C}/\text{cm}^2$	hydrophobic	SL200



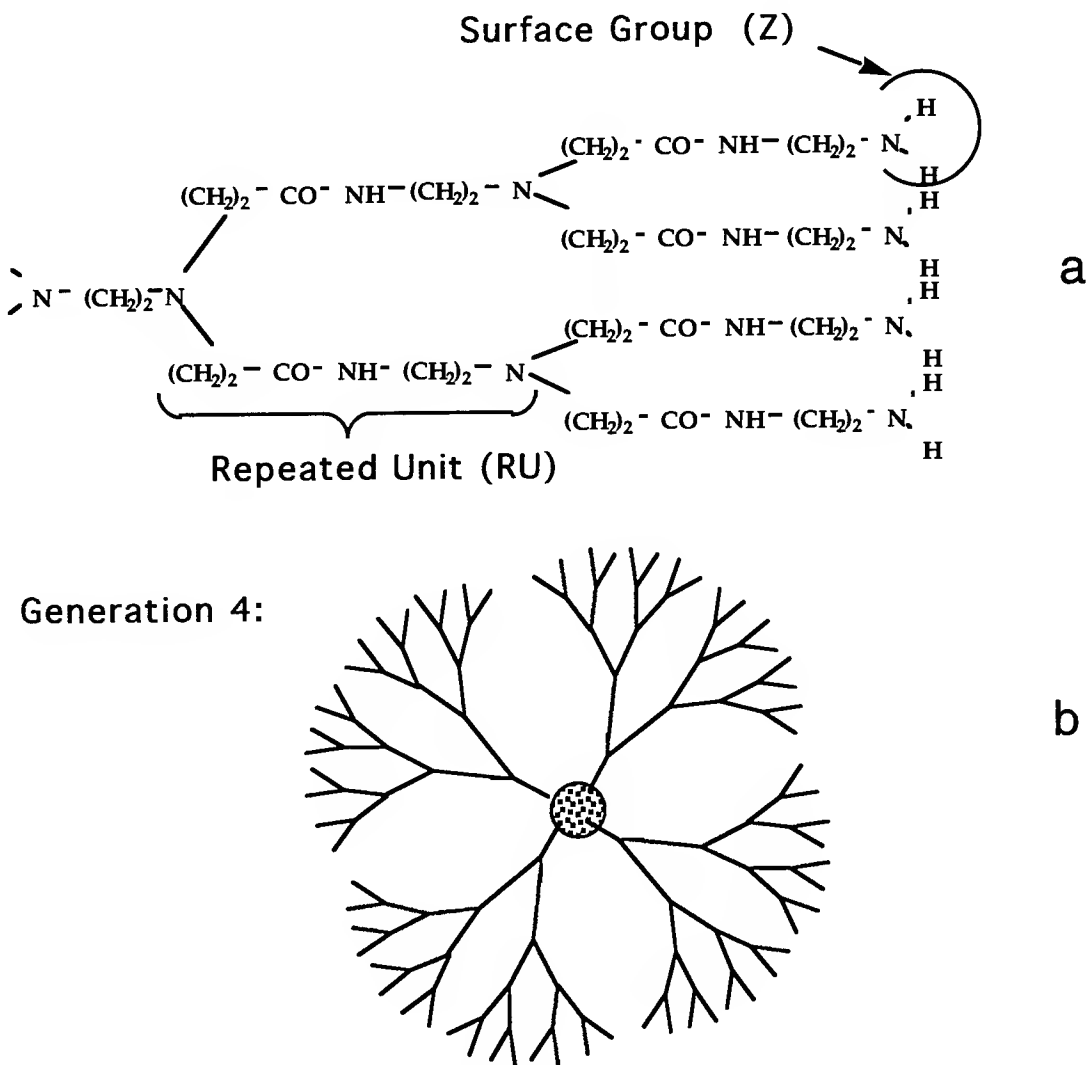


Figure 3. Chemical structure of dendrimers and a scheme of G4 dendrimer.

The positively charged surface is an amine terminated self-assembled monolayer (SAM) and the negatively charge substrate is a silicon oxide layer of a silicon wafer. An electrostatic layer-by-layer deposition technique was employed for the formation of the films from aqueous solution at neutral conditions ( $pH = 6.5$ ). At this pH both polyions possess some net charge as a result of dissociation (PSS) and protonization (PAA) of side groups. Solid substrates used were silicon wafers ((100) orientation, SAS) and float glass (Aldrich) modified by 3-aminopropyl triethoxysilane. Cleaning and modification of the substrates as well as formation and transfer of the monolayers onto the solid supports were performed using rigorous procedures.<sup>3</sup> All PSS/PAA films were prepared in a class 100 clean room and stored in a sealed containers.

Silanized substrates were protonated in a silicon wafer holder with a water solution of 0.01 N HCl. The 0.01 N HCl was poured into the holder and allowed to react for 2 min. The HCl was then poured out and the substrates in the holder were flushed under running Milli-Q water. The wafer pieces were then rinsed with Milli-Q water individually and dried with dry nitrogen. The concentrations of the aqueous PSS and PAA solutions were 2 mg/ml. The protonated substrates were dipped in the PSS solution for the appropriate amount of time as designated below (from 1 second to 64 minutes). A set of substrates were dipped in the PSS solution for 64 minutes and were used for the dipping in the PAA solution for different times as designated below. This procedure for sample fabrication described in detail for PSS/PAA films was basically unchanged for other types of self-assembling films studied here.

Atomic force (AFM) and friction force (FFM) images of fabricated films in contact and non-contact (the "tapping") modes were obtained in air at ambient temperature with the Nanoscope IIIA - Dimension 3000 (Digital Instruments, Inc.) according to well-established procedures (Figure 4).<sup>13</sup> We observed that a combination of the tapping mode or contact mode scanning with tip modification allowed stable reproducible imaging of the soft monolayers without visible damage. Images were obtained on scales from 200 nm to 100  $\mu\text{m}$  but for further analysis we selected two most appropriate "standard" sizes of 2 x 2  $\mu\text{m}$  and 5 x 5  $\mu\text{m}$ . All microroughness values reported here were measured for 1  $\mu\text{m}$  x 1  $\mu\text{m}$  areas.

The AFM tips were chemically modified to introduce appropriate surface charge and avoid tip contamination by charged macromolecules from the specimens.<sup>13</sup> Tip radii were in the range 20 - 40 nm as estimated by scanning a standard specimen with tethered colloidal gold nanoparticles with known diameters according to the published procedure.<sup>13</sup> AFM images were obtained for several specimens prepared under identical conditions at different periods of time, and at several randomly selected film areas. All structural parameters discussed below were averaged over 7 - 10 independent measurements after image processing.

X-ray reflectivity measurements were performed on a Siemens D-5000 diffractometer equipped with a reflectometry stage. X-ray data were collected within 0 - 70° scattering angle

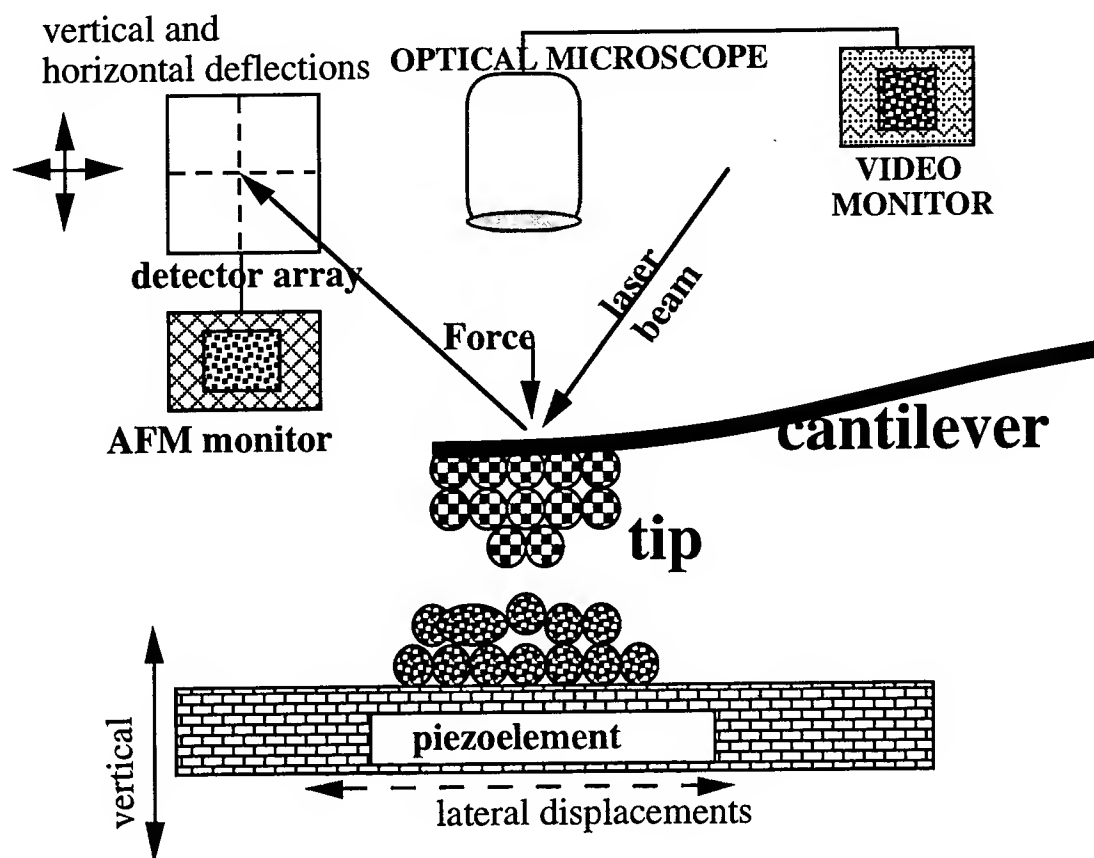


Figure 4. General scheme of AFM technique.

range with step of  $0.02^\circ$  using monochromatized  $\text{CuK}_\alpha$  radiation. Measurements within two angle intervals with different parameters (X-ray tube power and accumulation time) were rescaled to one angle interval. Simulation of X-ray curves was done by the REFSIM 1.0 program by direct computation of Fresnel reflectivity.<sup>14</sup> We used standard database densities and refractive indices for the substrates. Polymer film refractive indices were determined from database data for chemical elements in accordance with their chemical composition. For X-ray reflectivity simulations we used a double-layer model of surface structure of silicon substrates (silicon-silicon dioxide layers) with parameters determined independently for bare substrates. For polymer monolayers, we accepted homogeneous density distribution along the surface normal within a single molecular layer with Gaussian interfacial zones. Fitting parameters for polymer films were thickness, specific gravity, and roughness of polymer films.

All technical details of experimental procedures can be found in original publications.<sup>10-13</sup>

## Results and discussion

### *Polyion monolayer formation.*

Formation of self-assembled monolayers is monitored for PSS adsorbed on charged SAM surfaces and PAA on a PSS monolayer. In both cases, polyions are adsorbed on oppositely charged surfaces. Observations of PSS monolayers at various stages of electrostatic deposition reveal inhomogeneous self-assembly at the earliest stages of deposition. During the first several minutes, negatively-charged PSS macromolecules tend to adsorb on selected defect sites of positively charged SAM (scratches, microparticles, and edges) and form islands composed of PSS coils (Figure 5a). At this stage, electrostatic adsorption of PSS chains is predominant and equilibration of the surface structure is not achieved by the slow surface diffusion mechanism. Temporal variation of structural parameters of monolayers and bilayers are collected in Table 2.

Table 2. Structural parameters of adsorbed PSS monolayers and PAA/PSS bilayer.

Time, sec	height, PSS <sup>1</sup>	height, PSS <sup>2</sup>	height, PAA <sup>1</sup>	rms, PSS <sup>1</sup>	rms, PAA <sup>1</sup>
0	0	0.0	0	0.23	0.33
10	1.5	1.3		0.17	
20		1.2		2.04	
45	4.0	3.5	1.5	0.17	0.42
120	1.5	2.4	1.4	0.74	
300	1.6	1.7	0.9	0.64	0.32
600	1.0	1.6	1.0	0.16	0.23
1800	2.1	1.2		0.24	
2100			0.7		0.24
3000	1.5	1.2			
3840	1.0	2.0	0.9	0.26	0.22

All data are in nm; <sup>1</sup> data from AFM measurements, <sup>2</sup> data from ellipsometry; X-ray thickness for complete PSS monolayer and complete PAA/PSS bilayer is 1.5 nm and 2.6 nm, respectively.

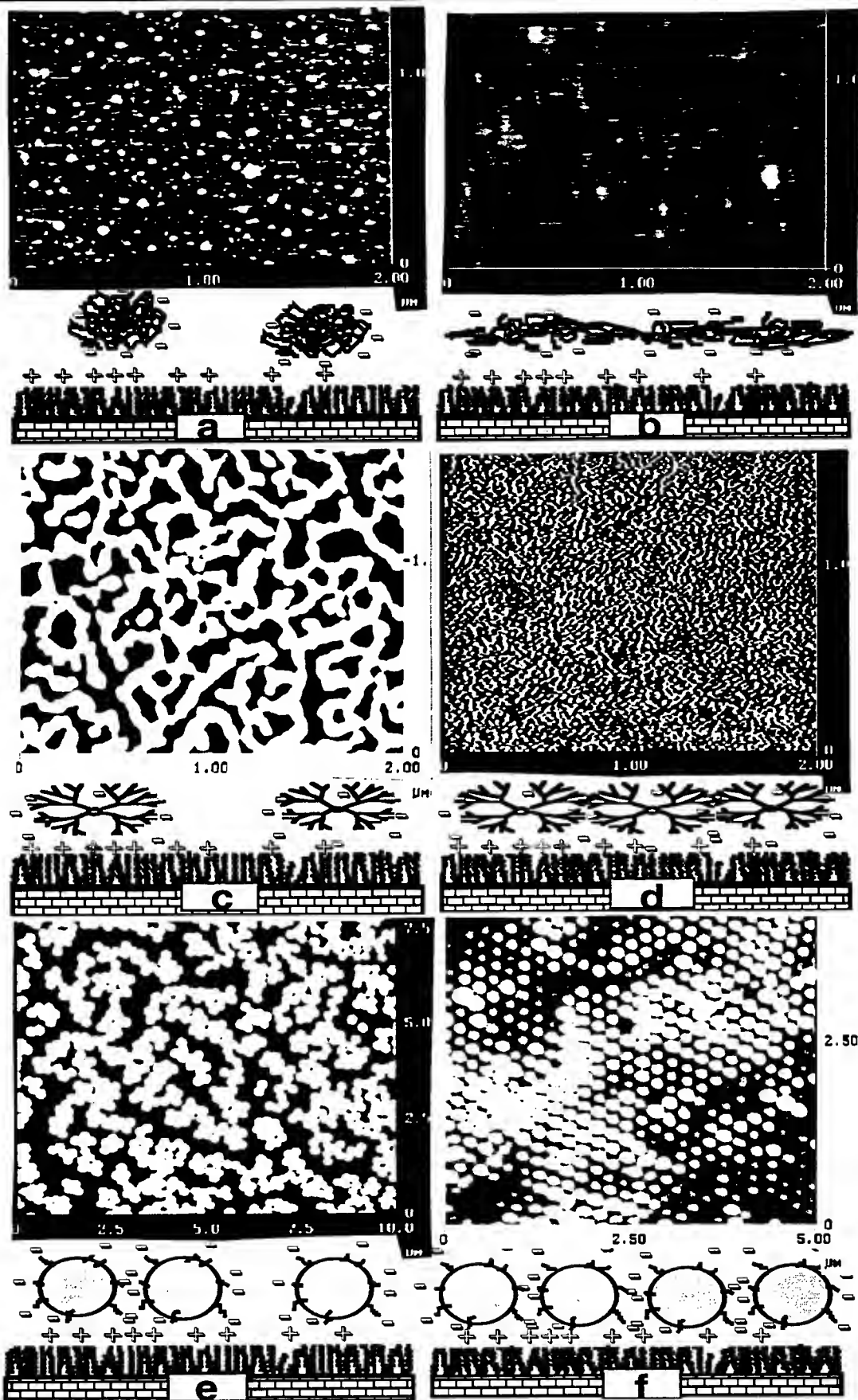


Figure 5. Topographical images of surface morphology and models of monolayer packing for amorphous polyions (a, b), dendrimers (c, d), and latex nanoparticles (e, f) at early stage (a, c, e) and in dense state (b, d, f).

Only longer deposition times result in an equilibration of polymer layers and formation of homogeneous thin PSS layer composed of highly flattened macromolecular chains (Figure 5b). The monolayer thickness is between 1.0 - 1.5 nm with a microroughness of about 0.2 nm. Therefore, the chains form a very thin molecular layer with a thickness of not more than 2 - 3 molecular cross-sections. Self-assembly of a second PAA layer on top of a PSS monolayer follows similar tendencies resulting in the formation of homogeneous PAA/PSS bilayers with an overall thickness of 1.7 - 2.5 nm. This bilayer is stable and cannot be damaged by the AFM tip.

#### *Dendritic monolayers*

At the initial stages of formation, isolated islands and network microstructures are detected for various Starburst<sup>15</sup> dendrimers (Figure 3). All even generations of dendrimers are observed to form homogeneous, compact monolayers on a silicon surface (Figure 5c, 5d). X-ray reflectivity (Figure 6) allows independent measurements of the average thickness of dendritic monolayers. As observed, the thickness of a single monolayer depends upon generation increasing with molecular weight: 1.8 nm (G4), 2.8 nm (G6), and 5.6 nm (G10) (Figure 7).

The average thickness of a molecular layer is much smaller than the diameter of ideal spherical dendritic macromolecules. The model of molecular ordering of dendrimer films assumes compressed dendritic macromolecules of oblate shape with an axial ratio in the range of 1 : 3 to 1 : 5 depending upon generation (Figure 5d). A tendency to higher spreading of high generation dendrimers observed here corresponds to the surface behavior predicted by molecular dynamic simulations.<sup>16</sup> The high interaction strength between "sticky" surface groups along with short range Van der Waals forces and long range capillary forces are considered to be responsible for formation of compact monolayer structures. Strong interactions between oppositely charged groups of dendritic macromolecules from adjacent molecular layers (such as ionic binding and formation of multiplets) may be responsible for compression of the soft architecture of dendritic macromolecules within self-assembled films as proposed by molecular computer modeling.

#### *Latex monolayers.*

We observed various stages of formation of a latex monolayer on oppositely charged substrates.<sup>17</sup> Incomplete monolayers represented by clusters composed of tens-hundreds

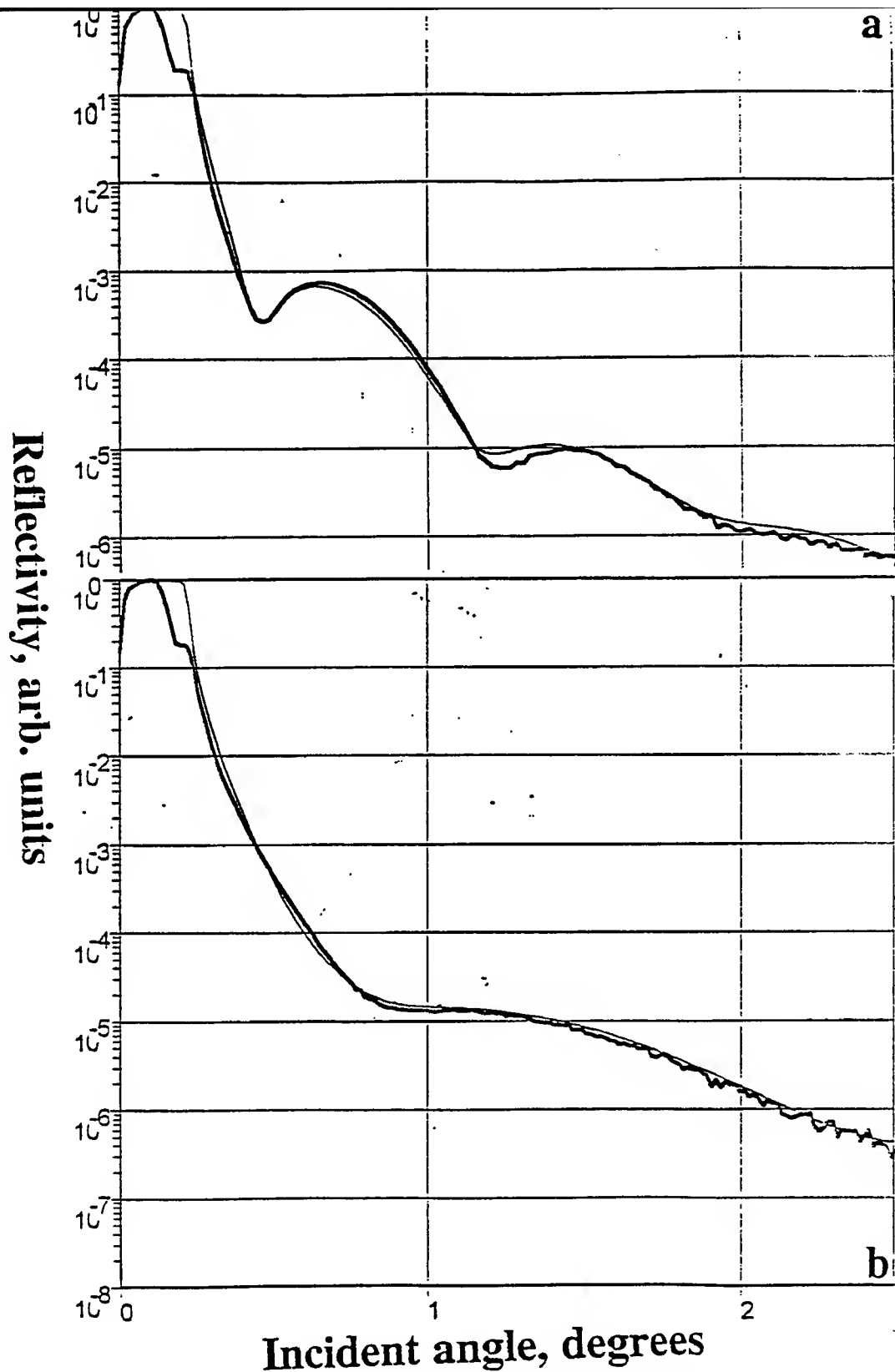


Figure 6. X-ray reflectivity experimental data (thick lines) and simulated curves (thin lines) for G6 (a) and G10 (b) monolayer films. Simulations curved were obtained for film thicknesses of 2.8 and 5.6 nm, roughness of 1.2 and 1.8 nm, and specific gravities of 1.2 and 1.3 g/cm<sup>3</sup> for G6 and G10 dendrimers, respectively.

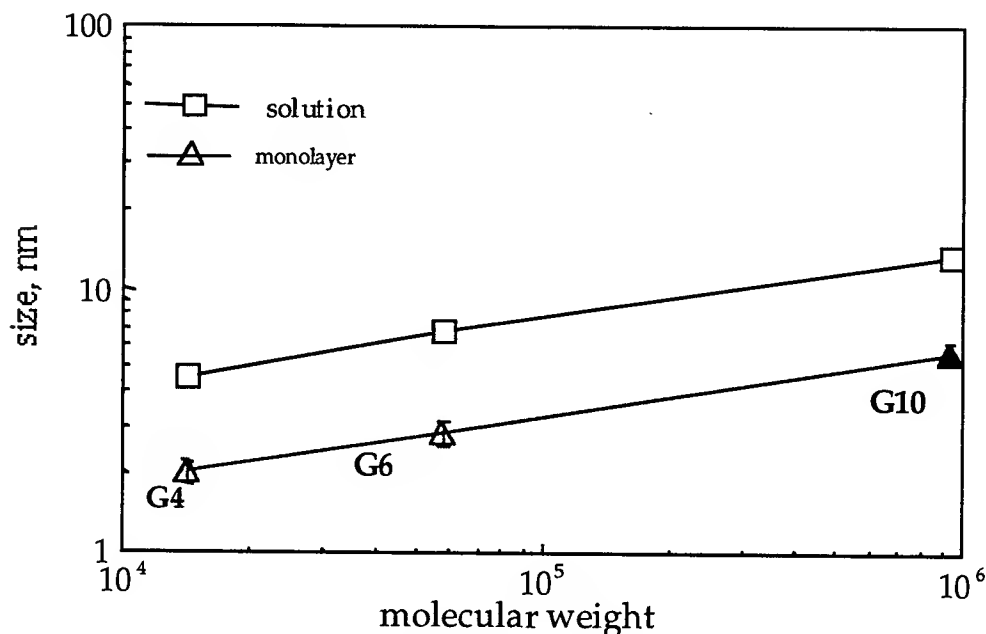


Figure 7. Spatial dimensions of dendrimers

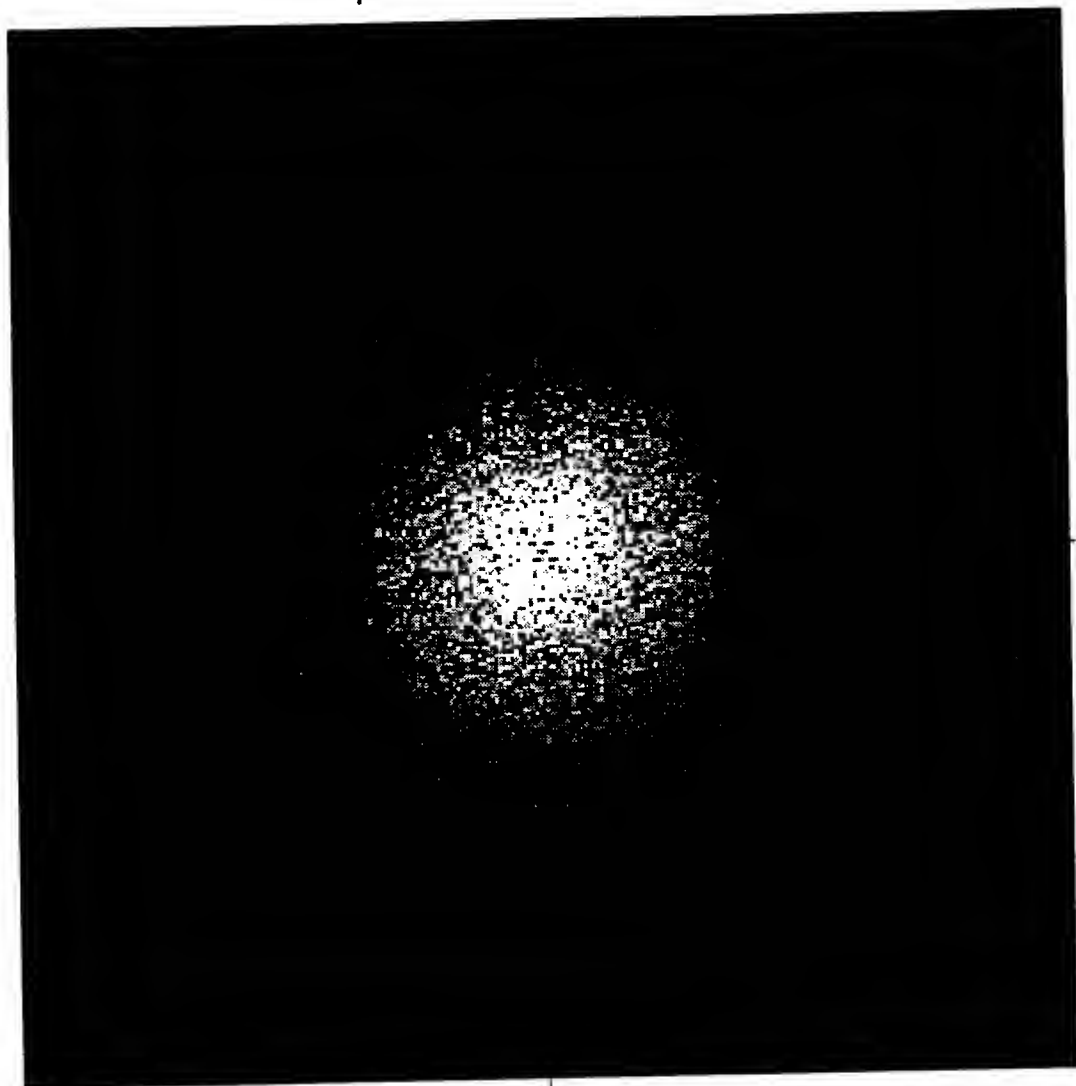
particles were formed for unassisted electrostatic adsorption (Figure 5e). These monolayers are composed of clusters of several dense packed latex nanoparticles randomly distributed over the surface. Maximum surface coverage can reach 70% for 20-nm latex nanoparticles.

Most of latex nanoparticles are able to form monolayers with short-range local ordering. These monolayers possess liquid type lateral packing of the nanospheres with a positional correlation expanded only over the nearest neighbors (Figure 8). Fourier analysis shows weak diffuse halo which corresponds to hexagonal packing and short-range ordering expanded over 4 - 5 coordination spheres (Figure 8). Larger latex nanoparticles with a narrow size distribution can form 2D lattices with long-range ordering within monolayers.<sup>11</sup> Obviously, strong tethering of charged nanoparticles to surfaces prevents their surface diffusion and rearrangements required for the formation of perfect lateral ordering. Formation of smooth monolayers composed of melted material is observed by thermal treatment at high temperature.

To overcome the strong repulsive forces among charged particles which prevents them from forming complete monolayers, we tested force assisted electrostatic self-assembly approach. We used additional capillary forces within the meniscus to form ordered monolayers during slow controlled pulling out of solution (modified dip-coating) and shear



# Spectrum 2D



$0.078 \mu\text{m}$  DC  $0.078 \mu\text{m}$

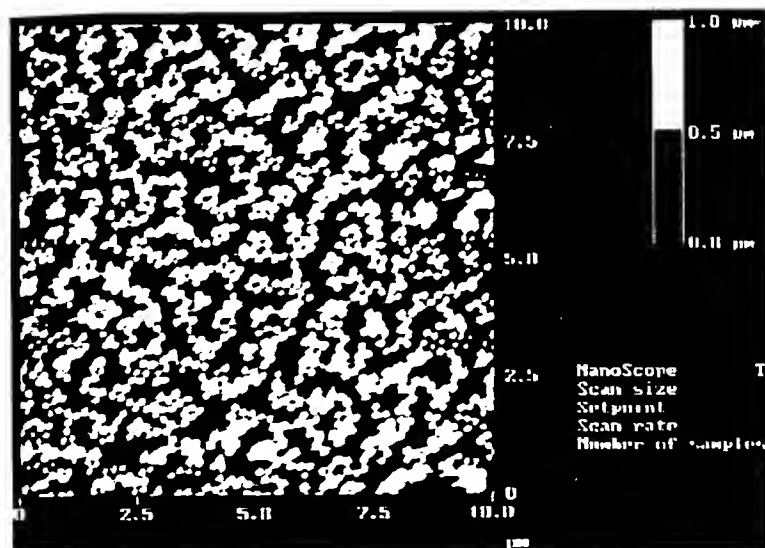


Figure 8. Fourier-transformation of latex monolayer with short-range in-plane ordering (insert).

# Spectrum 2D

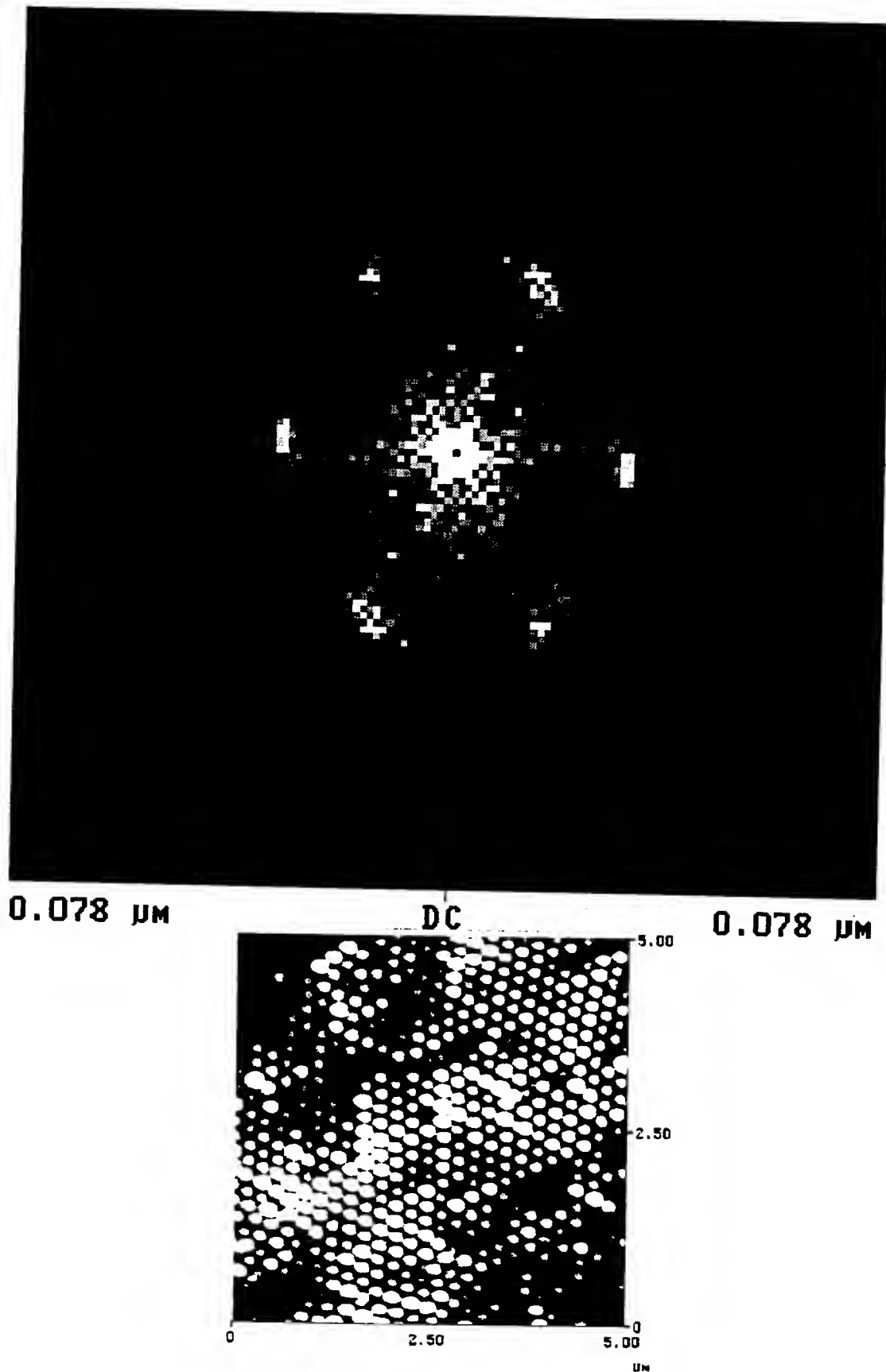


Figure 9. Fourier-transformation of latex monolayer with long-range in-plane ordering (insert).

flow resulting from spin-coating.<sup>17</sup> We observed that force assisted self-assembly of the same latexes can produce dense monolayer films with long-range in-plane ordering (Figure 5f). Lateral sizes of very well ordered monolayers are in the range of tens - hundreds of micrometers and usually six-fold symmetry of in-plane ordering is revealed by Fourier-analysis of monolayer films (Figure 9).

#### *Multilayer films from different polymeric materials*

Multilayer PSS/PAA films were fabricated and extensively studied earlier.<sup>18</sup> Virtually linear growth of film thickness versus number of deposition cycles was observed with an average increment of multilayer growth close to 1 nm (Figure 10). This value is close to monolayer thickness and similar to other amorphous polyions (see data for PVC/PAA film in Figure 10).<sup>18</sup>

Layer-by-layer deposition of oppositely charged dendrimers in combination  $G_n/G_{n-1/2}$  results in the formation of films with homogeneous surface morphology for a limited number of layers. A variation of the multilayer film thickness with the number of deposited layers,  $d(x)$ , is close to a linear with increment per layer in the range of  $2.8 \pm 0.3$  nm for G4/3.5 films and  $3.8 \pm 0.6$  nm for G10/9.5 film (see data for G4/3.5 in Figure 10). The small increment of the film thickness per molecular layer for multilayer films indicates that the dendritic macromolecules are indeed very soft and do not preserve their shape in the condensed state at interfaces similarly to monolayers. However, the average thickness of a molecular layer within the multilayer films is still two - three times higher for dendrimers than for amorphous polyions with comparable molecular weight despite their highly compressed state (Figure 10).

Typical multilayer growth pattern (thickness versus number of deposited layers) for CML20/AL20 latexes with 20 nm diameter is shown in Figure 7. The average increment for the first five layers is about 15 nm that corresponds to centered cubic packing of spheres. However, for  $n > 5$  the increment decreases to 7.3 nm per layer due to incomplete filling of following layers during the film growth. Surface roughness gradually increases for the first five layers and reaches a constant value of  $22 \pm 2$  nm for thicker films. This suggests that

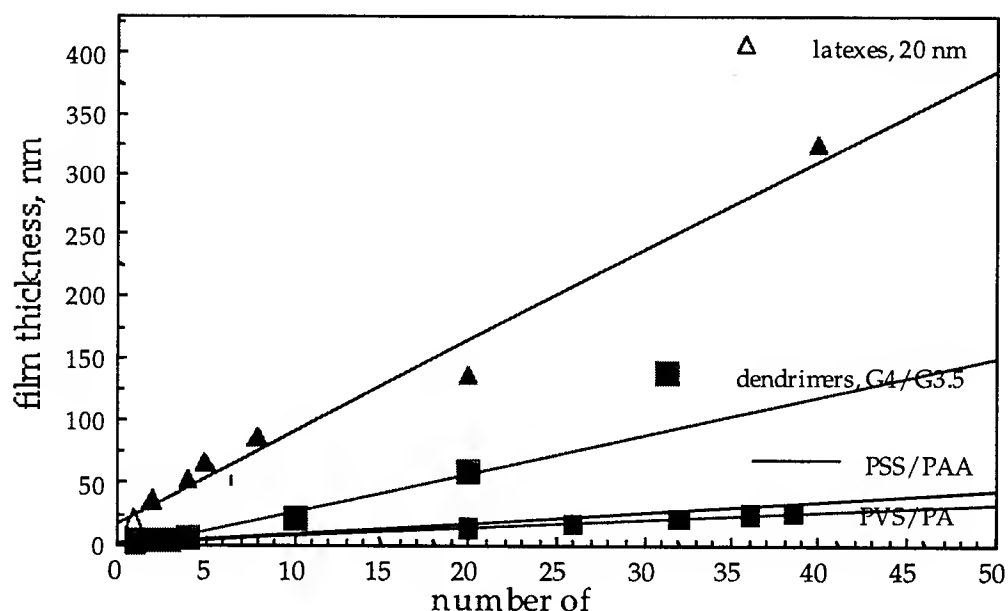


Figure 10. Film thickness growth for different polymer pairs.

some equilibrium growth process is established during this stage with roughly two layers being “under construction” during each deposition cycle.

Variation of pH of latex solution and activation of substrate with different pHs allow controlling a fine balance of interparticle and particles-substrates interactions due to change surface charge of ionizable surface groups. We tested various combinations of repulsive-attraction interaction strengths for PS latexes. We observed that weak repulsive interaction among nanoparticles (e. g., pH = 7 - 8 for amine terminated particles with pK = 9) combined with strong attraction of highly charged substrate (a bare silicon activated by solution with pH = 9) and slow pulling through the meniscus produces perfect monolayers and bilayers expanded over a surface area of a fraction of a millimeter across (Figure 11).

Fourier image shows a net of reflexes similar to a single-crystal pattern in reciprocal space which corresponds to long-range hexagonal packing of nanoparticles with interplanar distance (110) of 185 nm (Figure 11). This example represents the first truly macroscopically ordered monolayer of charged latex nanoparticles obtained by force assisted electrostatic deposition at mild ionic conditions.

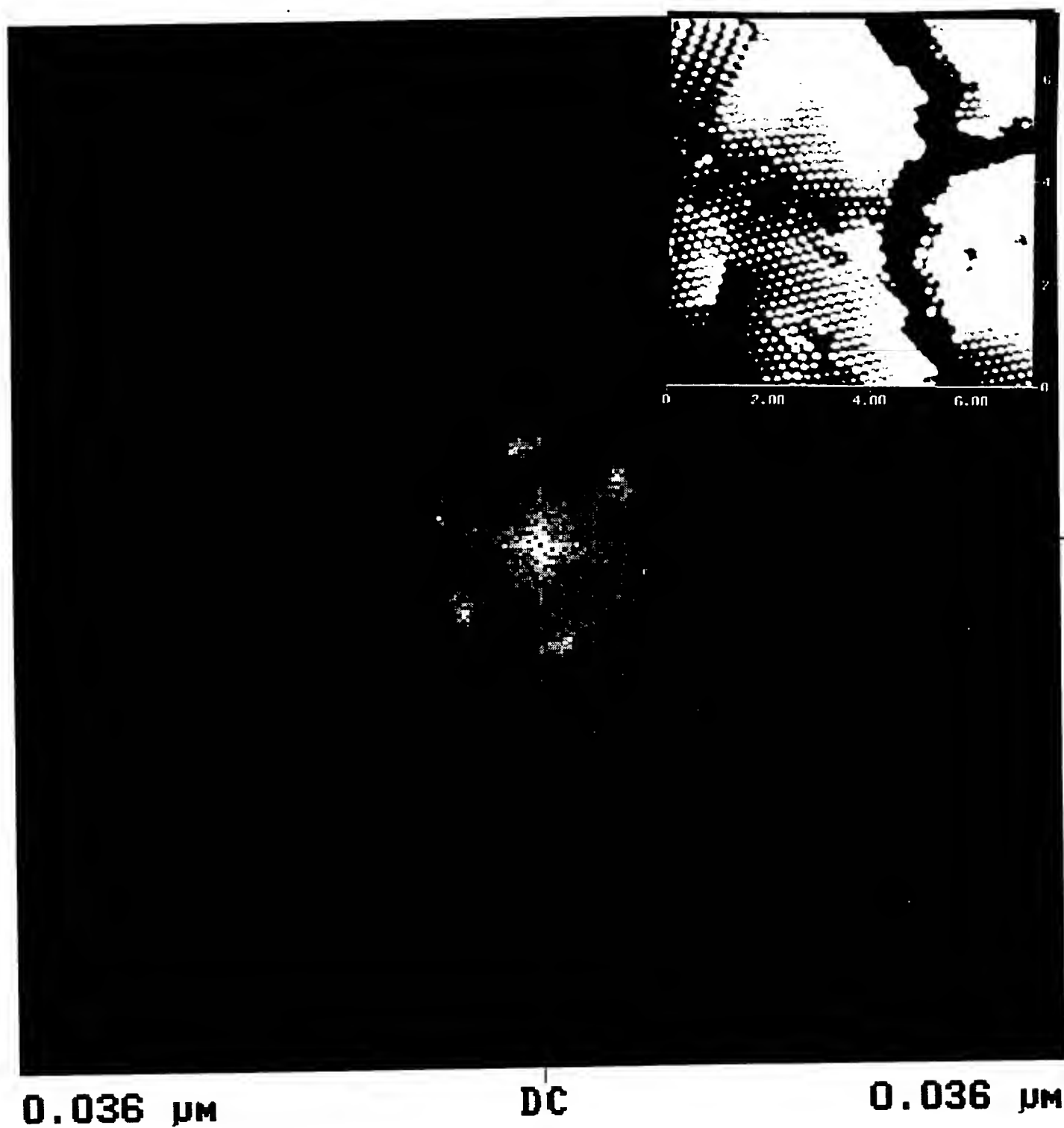


Figure 11. AFM image of latex nanoparticles organized in a perfect monolayer (a defected area is selected to underline smooth part) and corresponding Fourier-transformation.

*Multilayer films from organic-inorganic nanoparticles.*

Inorganic composite nanoparticles were tested as a prospective component for multilayer organic-inorganic films to enhance gradient of refractive index. Nanoparticles

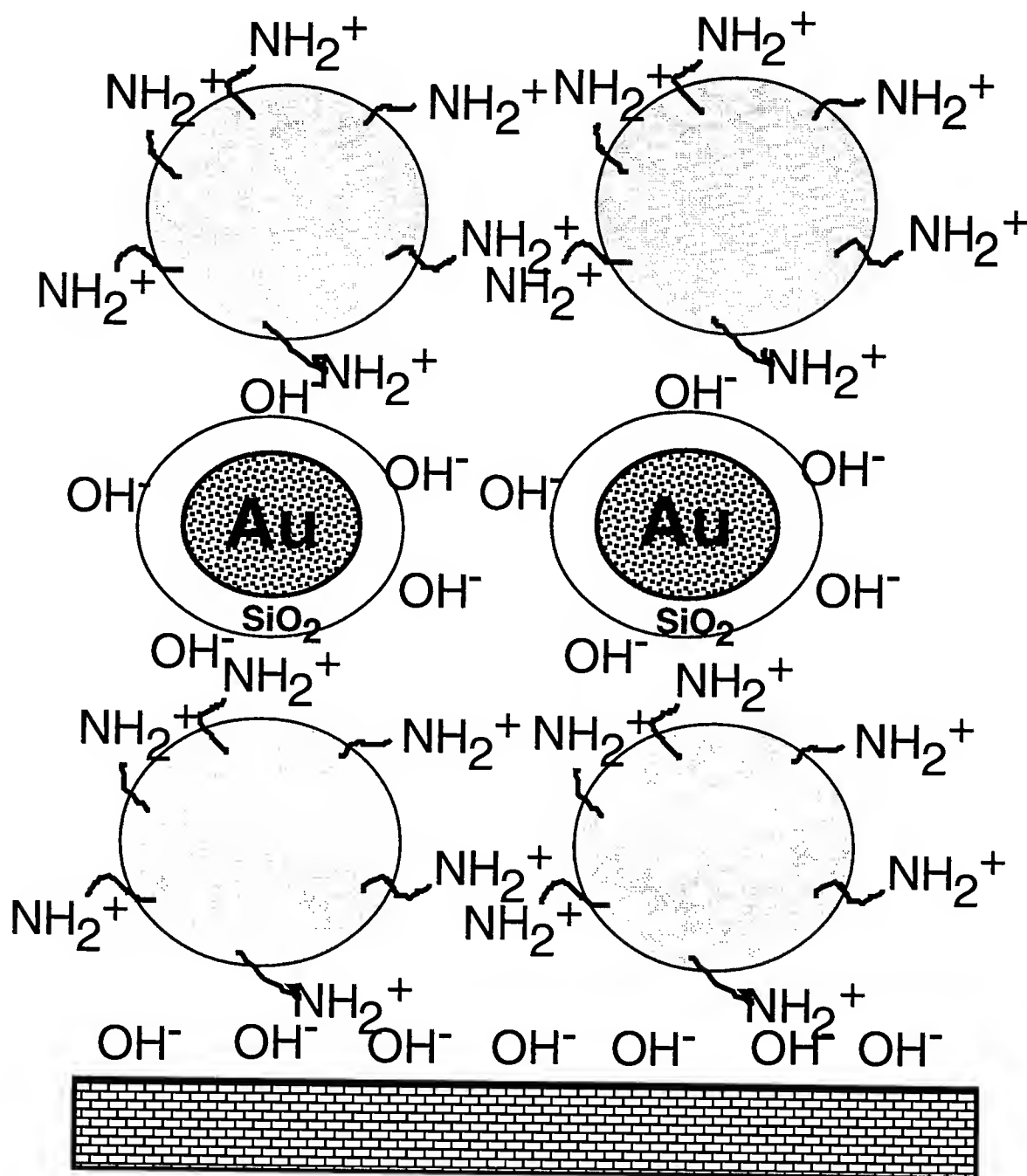


Figure 12. Organic-inorganic multilayers from latex/gold nanoparticles

tested are gold-core spheres with silicon oxide shell of an average diameter of 50 nm obtained from Melbourne University (Figure 12). Surface of these particles is terminated with silanol groups SiOH which is negatively charged at neutral pH. These nanoparticles can be used for monolayer fabrication on positively charged surfaces (amine SAMs) or as counterpart for positively charged amidine latex nanoparticles of 40-50 nm in a diameter. Our first attempts of monolayer fabrication showed promising results. Monolayers with dense packing of gold nanoparticles can be formed on amine terminated SAMs (Figure 13). Further studies are required for utilization of these inorganic nanoparticles for fabrication of multilayer films.

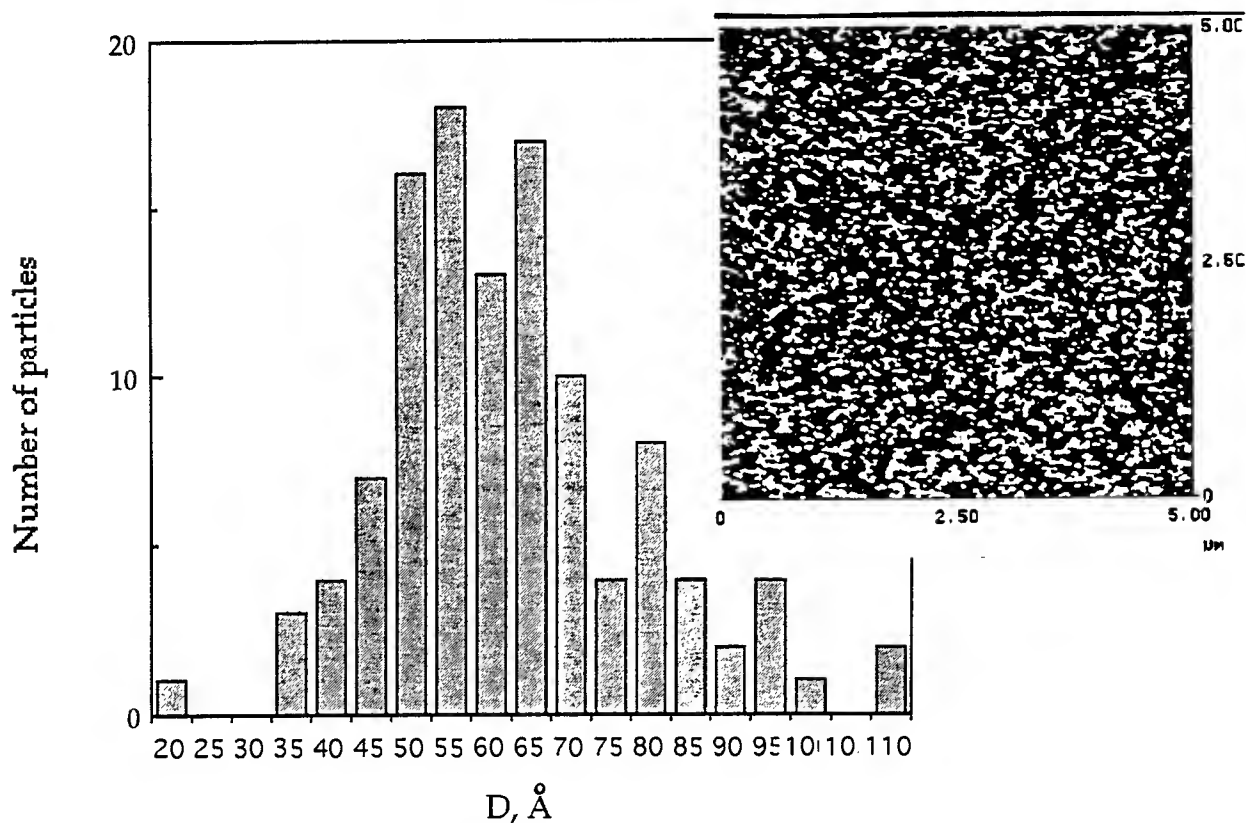


Figure 13. AFM image of monolayers of composite gold nanoparticles obtained by electrostatic deposition and histogram of height distribution.

## Conclusions and prospectives

From a comparison of the growth modes of multilayer films composed of amorphous polyionic materials and dendrimer/latex nanoparticles, we can conclude that replacing unstructured coiled macromolecular chains with organized dendritic supramolecules or “bulk” nanoparticles results in a significant increase in the growth increment. Correspondingly,

internal periodicity of these supramolecular multilayer self assembled films is about an order of magnitude higher than for conventional films. This type of structural organization is unachievable for amorphous polyelectrolytes with random spatial distribution of “sticky” groups. Obviously, the routine proposed may be used for formation of supramolecular films with mesoscale periodicity and intriguing optical properties.

General observation can be made that *force assisted electrostatic self assembly* of charged nanoparticles under dipping or spinning conditions may result in fabrication of relatively large uniform layers (a fraction of a millimeter across) with long-range positional and orientational ordering of nanoparticles within monolayers. However, only a fine balance of interparticle (particle-particle and particle-substrate) interactions and external forces within a narrow window may be useful for fabrication of perfect films. We present the first truly macroscopically ordered monolayer of charged latex nanoparticles obtained by force assisted electrostatic deposition. Lateral sizes of this monolayer are extended to a fraction of a millimeter.

#### Prospectives and trends

Several apparent problems and possible prospective trends can be deduced from our introductory study. These problems and trends should be addressed in more extensive project on a long term support basis.

#### *Problems to address are:*

- fabrication of complete, compact, and uniform molecular layers from nanoparticles with strong repulsive interactions, finding a right balance of interparticle and particle-substrate interactions and surface mobility
- stable growth of multilayer films with a large number of layers and prevention of in-plane phase separation at certain level when 3D bulk microphase structure becomes the most favorable
- “freezing” internal gradient of component composition along the surface normal for an indefinite long time

#### *Prospective directions to go and general trends could be defined as follow:*

- replacing self-assembly with “force assisted” self-assembly by addition active external gradient to overcome long-range repulsive interactions among mesoscale nanoparticles:



capillary forces (dip-coating), shearing flow (spin-coating), and electrostatic interactions (substrate potential variation and degree of ionization of functional groups)

- replacing pure polymer-polymer systems with organic-inorganic systems to enhance gradient of chemical composition and refractive index; appropriate inorganic nanoparticles “compatible” with existing polymer components should be thought
- testing a new generation of supramolecular assemblies which should include dendritic macromolecules with rigid architecture and functional surface groups able to form mesomorphic phases and zwitterionic/dye containing latex nanoparticles with functional surface groups

### **Acknowledgments**

This work is supported by US Air Force Office for Scientific Research, Contract F49620-93-C-0063. The PI thanks for contributions of the members of his research group: V. Bliznyuk, J. Hazel, D. Visser, J. Wu, S. Mirmiran, and J. Kalsi. My great thanks A. Campbell, T. Bunning, and W. Adams (WPAFB) for fruitful collaboration, J. H. Wendorff (University of Marburg) for X-ray and computer facilities provided, and P. Mulvaney (Melbourne University) for composite gold nanoparticles supply.

## References

1. Boivin, G.; St.-Germain, D. *Appl. Optics*, **1987**, 26, 4209; Bovard, B. *Appl. Optics*, **1993**, 32, 5427.
2. K. Komvopoulos, *Wear*, **1996**, 200, 305; Bhushan, B.; Israelachvili, J.; Landman, U. *Nature*, **1995**, 374, 607; Bhushan, B.; Kulkarni, A. V.; Koinkar, V. N.; Boehm, M.; Odoni, L.; Martelet, C.; Belin, M. *Langmuir*, **1995**, 11, 3189; Tsukruk, V. V.; Bliznyuk, V. N.; Hazel, J.; Visser, D.; Everson, M. P. *Langmuir* **1996**, 12, 4840; Tsukruk, V. V.; Everson, M. P.; Lander, L. M.; Brittain, W. J. *Langmuir* **1996**, 12, 3905.
3. Ulman, A. "Introduction to Ultrathin Organic Films", Acad. Pres.: San Diego, **1991**; Tredgold, R. *Order in Thin Organic Films*, Cambridge University Press: Cambridge, **1994**; Roberts, G. *Adv. Phys.* **1985**, 34, 475.
4. Tsukruk, V. V. *Progress in Polymer Science*, **1997**, 22, 247.
5. Fendler, J. H.; Meldrum, F. C. *Adv. Materials*, **1995**, 7, 607
6. Tsukruk, V. V.; Wendorff, J. H. *Trends in Polymer Science*, **1995**, 3, 82.
7. Iler, R. K. *J. Colloid and Interface Sci.* **1966**, 21, 569; Decher, G.; Hong, J.-D. *Makromol. Chem., Macromol. Symp.*, **1991**, 46, 321;
8. Decher, G.; Lvov, Yu.; Schmitt, J. *Thin Solid Films*, **1994**, 244, 772; Cooper, T. M.; Campbell, A. L.; Crane, R. L. *Langmuir* **1995**, 11, 2713; Ferreira, J. H. Cheung, M. Rubner, *Thin Solid Films*, 244, 806, **1994**; Lvov, Yu.; Ariga, K.; Kunitake, T. *Chem. Lett.* **1994**, 2323; Schmitt, T.; Grunewald, T.; Decher, G.; Pershan, P.; Kjaer, K.; Losche, M. *Macromolecules* **1993**, 26, 7058; Kellogg, G. J.; Mayes, A. M.; Stockton, W. B.; Ferreira, M.; M. F. Rubner; Satija, S. K. *Langmuir*, **1996**, 12, 5109.
9. V. V. Tsukruk, D. Janietz, *Langmuir*, 12, 2825, **1996**.
10. V. V. Tsukruk, V. N. Bliznyuk, D. W. Visser, A. L. Campbell, T. Bunning, W. W. Adams, **1996**, *Macromolecules*, submitted
11. V. N. Bliznyuk, V. V. Tsukruk, *Polymer Prepr.*, **1997**, 38 (1), 693.
12. V. V. Tsukruk, F. Rinderspacher, V. N. Bliznyuk, *Langmuir*, **1997**, accepted
13. Tsukruk, V. V.; Reneker, D. H. *Polymer*, **1995**, 36, 1791; V. N. Bliznyuk, J. Hazel, J. Wu, V. V. Tsukruk, in: *Scanning Probe Microscopy in Polymers*, Eds. V. V. Tsukruk, B. Ratner, ACS Symposium Series, 1997.
14. REFSIM 1.0, Siemens AG, Karlsruhe, 7500, Germany, 1994; Russell, T. P. *Materials Sci. Rep.*, **1990**, 5, 171; Foster, M. *Crit. Rev. Anal. Chem.*, **1993**, 24, 179; Tsukruk, V. V.; Shilov, V. V. *Structure of Polymeric Liquid Crystals*, Kiev, Naukova

- Dumka, 1990; Lipatov, Yu. S.; Shilov, V. V.; Gomza, Yu. P. *X-ray Analysis of Polymer Systems*, Kiev, Naukova Dumka, 1982; Hosemann, R.; Bagchi S. N. *Direct Analysis of Diffraction by Matter*, North Holland Publ. Co., Amsterdam, 1962; Tidswell, J. M.; Rabedeau, T. A.; Pershan, P.; Folkers, J. P.; Baker, M. V.; Whitesides, G. M. *Phys. Rev.*, **1991**, B44, 10869.
- 15.** Tomalia, D. A. *Adv. Materials*, **1994**, 6, 529; Jansen, J. F.; de Brabander-van den Berg, E. M.; Meijer, E. W. *Science*, **1994**, 266, 1226; Percec, V.; Kawasumi, M. *Macromolecules*, **1994**, 27, 4441; Frey, H.; Lorenz, K.; Holter, D.; Mulhaupt, R. *Polym. Prepr.*, **1996**, 37(1), 758; Frechet, J. M. *Science*, **1994**, 263, 1711. Sheiko, S. S.; Eckert, G.; Ignateva, G.; Musafarov, A. M.; Spickermann, J.; Rader, H. J.; Moller, M. *Makromol. Rapid Commun.*, **1996**, 17, 283.
- 16.** Mansfield, M. L. *Polymer*, **1996**, 37, 3835.
- 17.** N.D.Denkov, O.D.Velev, P.A.Krachevsky, I.B.Ivanov, H.Yoshimura, K.Nagayama, *Langmuir*, 8, 3183 (**1992**); A.S.Dimitrov, K.Nagayama, *Langmuir*, 12, 1303 (**1996**). J. Li, D. J. Meier, *Polymer Preprints*, 37 (2), 591 (**1996**); C.A.Johnson, A.M.Lenhoff, *J. Colloid and Interface Sci.*, 179, 587 (**1996**)
- 18.** Lvov, Yu. M.; Decher, G. *Crystallography Reports* **1994**, 39, 628

Distributed Control of Nonlinear Flexible Beams and Plates  
with Mechanical and Temperature Excitations

H. S. Tzou  
Professor  
Department of Mechanical Engineering

University of Kentucky  
Lexington, KY

Final Report for:  
Summer Research Extension Program  
Wright Laboratory

Sponsored by:  
Air Force Office of Scientific Research  
Bolling Air Force Base, DC

and

Wright Laboratory

February 1997

# DISTRIBUTED CONTROL OF NONLINEAR FLEXIBLE BEAMS AND PLATES WITH MECHANICAL AND TEMPERATURE EXCITATIONS

H. S. Tzou

Professor

Department of Mechanical Engineering  
University of Kentucky  
Lexington, KY 40506-0108

## Abstract

Beam and plate-type components are widely used in many aerospace structures. Imposed shape changes and surface control of flexible beams and plates could offer many aerodynamic advantages in flight maneuverability and precision control. Deformed shapes and surfaces often involve nonlinear deformations. Studies of control of nonlinear behavior related to the deformed surfaces and shape changes would provide detailed information in future controlled surface design and implementation. This research is concerned with the control effectiveness of nonlinearly deformed beams and plates based on the smart structures technology. Piezoelectric materials are widely used as sensors and actuators in sensing, actuation, and control of smart structures and structronic systems. Control effectiveness of piezoelectric laminated nonlinear flexible beams and plates subjected to mechanical and temperature excitations is investigated. It is assumed that the flexible beams and plates encounter the von Karman type geometrical nonlinearity. Thermoelectromechanical equations and boundary conditions including elastic, temperature, and piezoelectric couplings are formulated first, and analytical solutions derived next. Dynamics, active control of nonlinear flexible deflections, thermal deformations, and natural frequencies using distributed piezoelectric actuators are studied, and their nonlinear effects are evaluated.

# DISTRIBUTED CONTROL OF NONLINEAR FLEXIBLE BEAMS AND PLATES WITH MECHANICAL AND TEMPERATURE EXCITATIONS

H. S. Tzou

## INTRODUCTION

Recent development of smart (or intelligent) structures, structronic (structure-electronic) systems, and micro mechanical systems has demonstrated the versatilities of piezoelectric materials in both sensor, the *direct piezoelectric effect*, and actuator, the *converse piezoelectric effect*, applications (Tzou and Anderson, 1992; Tzou and Fukuda, 1992). Piezoelectrics are usually bonded (embedded or surface coupled) with elastic structures, serving as sensors and/or actuators for structural monitoring and control. These piezoelectric sensors and actuators can be further classified as "discrete" or "distributed" devices. The distributed sensors and actuators offer many advantages over the discrete devices, such as multiple modal controls, spatial filterings, spatially shaped modal controls, etc (Tzou, 1993). Accordingly, distributed piezoelectric sensors and actuators are widely used in various structural applications.

In recent years, many sophisticated analyses are performed and new engineering applications are explored. However, most of these studies were conducted based on the linear theories, i.e., linear elasticity and piezothermoelectricity. It is known that piezoelectric materials are rather nonlinear. In addition, large oscillation and/or flexibility of elastic structures can introduce large deflections of distributed sensors and actuators. Thus, the nonlinear characteristics of piezoelectric materials and laminated structures can be classified into two categories: 1) the geometrical nonlinearity and 2) the material nonlinearity. The former usually involves large deformations and the latter is associated with nonlinear material properties, e.g., hysteresis, temperature dependent material constants, etc (Tzou and Bao, 1996). Pai, et al. (1993) recently studied a nonlinear composite plate laminated with piezoelectric layers. Yu (1993) reviewed recent studies of linear and nonlinear theories of elastic and piezoelectric plates. Lalande et al. (1993) investigated the nonlinear deformation of a piezoelectric Moonie actuator based on a

simplified nonlinear beam theory. Sreeram et al. (1993) investigated a nonlinear hysteresis modeling of piezoceramic actuator. Librescu (1987) proposed a refined geometrical nonlinear theory of anisotropic laminated shells. Linear thermo-electromechanical behavior of distributed piezoelectric sensors and actuators were also recently studied (Tzou and Ye, 1994; Tzou and Howard, 1994). A theory on geometrical nonlinearity of piezothermoelastic shell laminates simultaneously exposed to mechanical, electric, and thermal fields has been recently proposed (Tzou and Bao, 1996). Tzou and Zhou (1995) investigated static and dynamic control of a circular plate with geometrical nonlinearity. This research is devoted to a study of distributed static and dynamic control of nonlinear flexible beams and rectangular plates with geometrical nonlinearity using distributed piezoelectric actuators.

Since flexible beams and plates (rectangular and square plates) are very common in aerospace structures, static and dynamic behaviors of piezothermoelastic laminated flexible beams and plates with initial large nonlinear deformations (the von Karman type geometrically nonlinear deformations) and subjected to mechanical, electric, and temperature excitations are investigated in this research. Active control effects on nonlinear static deflections and natural frequencies, including temperature variations, imposed by the piezoelectric actuators are investigated. Nonlinear beam and plate equations are derived first, followed by nonlinear static analysis and free vibration analysis including the effect of initial nonlinear deformations. Active control of nonlinear effects by the piezoelectric actuators are emphasized. Numerical examples are provided and simulation results discussed.

## CONSTITUTIVE EQUATIONS

This study focuses on the distributed control effectiveness of nonlinear deformations and dynamic frequencies, including mechanical, electric, and temperature effects. Since the fundamental governing equations are derived from the triclinic piezoelectric materials and thin anisotropic piezothermoelastic shells, original generic piezothermoelastic governing equations are briefly reviewed. The relations between the electric fields  $E_1$ ,  $E_2$ ,  $E_3$  and the electric potential  $\varphi$  in the curvilinear coordinate system are (Tzou and Zhong, 1993)

$$\begin{bmatrix} E_1 \\ E_2 \\ E_3 \end{bmatrix} = - \begin{bmatrix} f_{11} & 0 & 0 \\ 0 & f_{22} & 0 \\ 0 & 0 & f_{33} \end{bmatrix}^{-1} \begin{bmatrix} \partial/\partial\alpha_1 \\ \partial/\partial\alpha_2 \\ \partial/\partial\alpha_3 \end{bmatrix} \varphi. \quad (1)$$

The constitutive relations of the piezothermoelastic shell are governed by three equations: 1) a stress equation  $\{T\}$ , 2) an electric displacement equation  $\{D\}$ , and 3) a thermal entropy equation  $\mathcal{E}$  (Tzou and Ye, 1994).

$$\begin{cases} \{T\} = [c]\{S\} - [e]^t\{E\} - \{\lambda\}\theta, & (2) \end{cases}$$

$$\begin{cases} \{D\} = [e]\{S\} + [\epsilon]\{E\} + \{p\}\theta, & (3) \end{cases}$$

$$\begin{cases} \mathcal{E} = \{\lambda\}^t\{S\} + \{p\}^t\{E\} + \alpha_v \theta, & (4) \end{cases}$$

where  $\{T\}$ ,  $\{S\}$ ,  $\{E\}$  and  $\{D\}$  denote the stress, strain, electric field and electric displacement vectors, respectively;  $\mathcal{E}$  is the thermal entropy density;  $\theta$  is the temperature rise ( $\theta = \Theta - \Theta_0$  where  $\Theta$  is the absolute temperature and  $\Theta_0$  the temperature of natural state in which stresses and strains are zero);  $[c]$ ,  $[e]$ , and  $[\epsilon]$  denote the elastic stiffness coefficient, piezoelectric coefficient, and dielectric permittivity matrices, respectively;  $\{\lambda\}$  is the stress-temperature coefficient vector;  $\{p\}$  is the pyroelectric coefficient vector; and  $\alpha_v$  is a material constant ( $\alpha_v = \rho c_v / \Theta_0$  where  $c_v$  is the specific heat at a constant volume).  $[\cdot]^t$  and  $\{\cdot\}^t$  are matrix and vector transpose, respectively. For a generic anisotropic piezothermoelastic material, the stress and electric displacement equations become (Tzou and Bao, 1994)

$$\begin{bmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \\ T_5 \\ T_6 \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} & c_{14} & c_{15} & c_{16} \\ & c_{22} & c_{23} & c_{24} & c_{25} & c_{26} \\ & & c_{33} & c_{34} & c_{35} & c_{36} \\ & & & c_{44} & c_{45} & c_{46} \\ & \text{sym.} & & & c_{55} & c_{56} \\ & & & & & c_{66} \end{bmatrix} \begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \end{bmatrix} - \begin{bmatrix} e_{11} & e_{21} & e_{31} \\ e_{12} & e_{22} & e_{32} \\ e_{13} & e_{23} & e_{33} \\ e_{14} & e_{24} & e_{25} \\ e_{15} & e_{25} & e_{35} \\ e_{16} & e_{26} & e_{36} \end{bmatrix} \begin{bmatrix} E_1 \\ E_2 \\ E_3 \end{bmatrix} - \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \\ \lambda_6 \end{bmatrix} \theta, \quad (5)$$

$$\begin{bmatrix} D_1 \\ D_2 \\ D_3 \end{bmatrix} = \begin{bmatrix} e_{11} & e_{12} & e_{13} & e_{14} & e_{15} & e_{16} \\ e_{21} & e_{31} & e_{41} & e_{24} & e_{25} & e_{26} \\ e_{31} & e_{32} & e_{33} & e_{34} & e_{35} & e_{36} \end{bmatrix} \begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \epsilon_{33} \end{bmatrix} \begin{bmatrix} E_1 \\ E_2 \\ E_3 \end{bmatrix} + \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} \theta. \quad (6)$$

Note that if the effective axes of the piezothermoelastic material do not coincide with the geometrical axes, an orientation or transformation matrix, in directional cosines and sines,



needs to be defined (Tzou and Bao, 1995). The anisotropic piezothermoelastic material is used in deriving the generic piezothermoelastic shell system equations; simplifications to other simpler piezothermoelastic materials, e.g., polyvinylidene fluoride, piezoceramics, etc., are then explored.

## VON-KARMAN NONLINEARITY

It is assumed that the nonlinear characteristics are introduced by large deformations which can be introduced mechanically, electrically, and/or thermally. A generic nonlinear deflection  $U_i$  in the  $i$ -th direction can be expressed as a summation of a membrane displacement  $u_i(\alpha_1, \alpha_2, t)$  and a higher order nonlinear shear deformation effect represented by the summation of angular rotations  $\beta_{ij}(\alpha_1, \alpha_2, t)$  (Tzou and Bao, 1996):

$$U_i(\alpha_1, \alpha_2, \alpha_3, t) = u_i(\alpha_1, \alpha_2, t) + \sum_{j=1}^m \alpha_3^j \beta_{ij}(\alpha_1, \alpha_2, t), \quad i=1,2,3, \quad (7)$$

where  $u_1$ ,  $u_2$  and  $u_3$  are the mid-plane displacement components of the reference surface along the  $\alpha_1$ ,  $\alpha_2$ , and  $\alpha_3$  axes, respectively;  $\beta_{11}$  and  $\beta_{21}$  represent the rotational angles in the positive sense of the  $\alpha_1$  and  $\alpha_2$  axes, respectively; and  $\beta_{3j} = 0$ . This expression includes higher order nonlinear shear deformation effects. However, according to the Love-Kirchhoff thin shell assumptions and a linear displacement approximation (first order shear deformation theory), only the first term is kept in the equation, i.e.,  $m = 1$  (Tzou, 1993). The displacements and rotational angles are independent variables in thick shells. However, the rotational angles are dependent variables in thin shells, and they can be derived from the thin shell assumptions in which the transverse normal strain  $S_3$  is negligible and shear strains  $S_4$  and  $S_5$  are zeros. Based on the thin shell assumptions, the rotational angles  $\beta_1 = \beta_{11}$  and  $\beta_2 = \beta_{21}$  are derived from the transverse shear strain equations, i.e.,  $S_4 = 0$  and  $S_5 = 0$ .

$$\beta_1 = \frac{u_1}{R_1} - \frac{1}{A_1} \frac{\partial u_3}{\partial \alpha_1}, \quad \text{and} \quad \beta_2 = \frac{u_2}{R_2} - \frac{1}{A_2} \frac{\partial u_3}{\partial \alpha_2}. \quad (8a,b)$$

In general,  $\alpha_3/R_1 \ll 1$  and  $\alpha_3/R_2 \ll 1$ , thus, the ratios of the finite distance to the radius of curvature are negligible, i.e.,  $f_{11} \simeq A_1$  and  $f_{22} \simeq A_2$ . It is assumed that the piezothermoelastic shell experiences large deformations in three axial directions. However,

in general, the in-plane deflections are still much smaller than the transverse deflections. Thus, the nonlinear effects due to the in-plane large deflections are usually neglected, i.e., the von Karman-type assumptions (Palazotto and Dennis, 1992; Chia, 1980). The nonlinear strain-displacement relations of a thin shell with a large transverse deflection  $u_3$  include a linear effect, denoted by a superscript  $l$ , and a nonlinear effect, denoted by a superscript  $n$ , induced by the large deformation:

$$\begin{bmatrix} S_1 \\ S_2 \\ S_6 \end{bmatrix} = \begin{bmatrix} \bar{s}_1^0 \\ \bar{s}_2^0 \\ \bar{s}_6^0 \end{bmatrix} + \alpha_3 \begin{bmatrix} \bar{\kappa}_1 \\ \bar{\kappa}_2 \\ \bar{\kappa}_6 \end{bmatrix} = \left[ \begin{bmatrix} \bar{s}_1^0 \\ \bar{s}_2^0 \\ \bar{s}_6^0 \end{bmatrix}^l + \begin{bmatrix} \bar{s}_1^0 \\ \bar{s}_2^0 \\ \bar{s}_6^0 \end{bmatrix}^n \right] + \alpha_3 \begin{bmatrix} \bar{\kappa}_1 \\ \bar{\kappa}_2 \\ \bar{\kappa}_6 \end{bmatrix}, \quad (9)$$

where the subscripts 1 and 2 respectively denote two normal strains and 6 is the in-plane shear strain. Detailed membrane and bending strains are functions of displacements  $u_i$ 's.

1) Membrane strains:

$$\bar{s}_1^0 = \frac{1}{A_1} \frac{\partial u_1}{\partial \alpha_1} + \frac{u_2}{A_1 A_2} \frac{\partial A_1}{\partial \alpha_2} + \frac{u_3}{R_1} + \left[ \frac{1}{2} \left( \frac{\partial u_3}{\partial \alpha_1} \right)^2 \right], \quad (10)$$

$$\bar{s}_2^0 = \frac{1}{A_2} \frac{\partial u_2}{\partial \alpha_2} + \frac{u_1}{A_1 A_2} \frac{\partial A_2}{\partial \alpha_1} + \frac{u_3}{R_2} + \left[ \frac{1}{2} \left( \frac{\partial u_3}{\partial \alpha_2} \right)^2 \right], \quad (11)$$

$$\bar{s}_6^0 = \frac{1}{A_2} \frac{\partial u_1}{\partial \alpha_2} + \frac{1}{A_1} \frac{\partial u_2}{\partial \alpha_1} - \frac{u_1}{A_1 A_2} \frac{\partial A_1}{\partial \alpha_2} - \frac{u_2}{A_1 A_2} \frac{\partial A_2}{\partial \alpha_1} + \left[ \frac{1}{A_1 A_2} \frac{\partial u_3}{\partial \alpha_1} \frac{\partial u_3}{\partial \alpha_2} \right]; \quad (12)$$

2) Bending strains:

$$\bar{\kappa}_1 = \frac{1}{A_1} \frac{\partial}{\partial \alpha_1} \left[ \frac{u_1}{R_1} - \frac{1}{A_1} \frac{\partial u_3}{\partial \alpha_1} \right] + \frac{1}{A_1 A_2} \left[ \frac{u_2}{R_2} - \frac{1}{A_2} \frac{\partial u_3}{\partial \alpha_2} \right] \frac{\partial A_1}{\partial \alpha_2}, \quad (13)$$

$$\bar{\kappa}_2 = \frac{1}{A_2} \frac{\partial}{\partial \alpha_2} \left[ \frac{u_2}{R_2} - \frac{1}{A_2} \frac{\partial u_3}{\partial \alpha_2} \right] + \frac{1}{A_1 A_2} \left[ \frac{u_1}{R_1} - \frac{1}{A_1} \frac{\partial u_3}{\partial \alpha_1} \right] \frac{\partial A_2}{\partial \alpha_1}, \quad (14)$$

$$\begin{aligned} \bar{\kappa}_6 = & \frac{1}{A_2} \frac{\partial}{\partial \alpha_2} \left[ \frac{u_1}{R_1} - \frac{1}{A_1} \frac{\partial u_3}{\partial \alpha_1} \right] + \frac{1}{A_1} \frac{\partial}{\partial \alpha_1} \left[ \frac{u_2}{R_2} - \frac{1}{A_2} \frac{\partial u_3}{\partial \alpha_2} \right] \\ & - \frac{1}{A_1 A_2} \left[ \frac{u_1}{R_1} - \frac{1}{A_1} \frac{\partial u_3}{\partial \alpha_1} \right] \frac{\partial A_1}{\partial \alpha_2} - \frac{1}{A_1 A_2} \left[ \frac{u_2}{R_2} - \frac{1}{A_2} \frac{\partial u_3}{\partial \alpha_2} \right] \frac{\partial A_2}{\partial \alpha_1}, \end{aligned} \quad (15)$$

where  $\bar{s}_1^0$ ,  $\bar{s}_2^0$  and  $\bar{s}_6^0$  are the membrane strains and  $\bar{\kappa}_1$ ,  $\bar{\kappa}_2$  and  $\bar{\kappa}_6$  are the bending strains (the change of curvatures on the reference surface). Note that the quadratic terms (nonlinear terms) inside the brackets are contributed by the large deflection. Membrane force resultants  $N_{ij}$  and bending moments  $M_{ij}$  of the piezothermoelastic shell laminate can be derived based on the induced strains:

$$\begin{bmatrix} N_{11} \\ N_{22} \\ N_{12} \\ M_{11} \\ M_{22} \\ M_{12} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & A_{16} & B_{11} & B_{12} & B_{16} \\ A_{12} & A_{22} & A_{26} & B_{12} & B_{22} & B_{26} \\ A_{16} & A_{26} & A_{66} & B_{16} & B_{26} & B_{66} \\ B_{11} & B_{12} & B_{16} & D_{11} & D_{12} & D_{16} \\ B_{12} & B_{22} & B_{26} & D_{12} & D_{22} & D_{26} \\ B_{16} & B_{26} & B_{66} & D_{16} & D_{26} & D_{66} \end{bmatrix} \begin{bmatrix} \bar{s}_1^e \\ \bar{s}_2^e \\ \bar{s}_6^e \\ \bar{\kappa}_1 \\ \bar{\kappa}_2 \\ \bar{\kappa}_6 \end{bmatrix} - \begin{bmatrix} N_{11}^e \\ N_{22}^e \\ N_{12}^e \\ M_{11}^e \\ M_{22}^e \\ M_{12}^e \end{bmatrix} - \begin{bmatrix} N_{11}^\theta \\ N_{22}^\theta \\ N_{12}^\theta \\ M_{11}^\theta \\ M_{22}^\theta \\ M_{12}^\theta \end{bmatrix}. \quad (16)$$

It is observed that there are three components, i.e., mechanical, electric, and temperature, in the force/moment expressions. Superscripts  $e$  and  $\theta$  respectively denote the electric and temperature components. The membrane strains and bending strains are coupled by the coupling stiffness coefficients  $B_{ij}$  in elastic force/moment resultants.  $N_{ij}^e$  and  $N_{ij}^\theta$  are the electric and temperature induced forces;  $M_{ij}^e$  and  $M_{ij}^\theta$  are the electric and temperature induced moments, respectively. In actuator applications, these electric forces and moments are used to control shell's static and dynamic characteristics.

## NONLINEAR PIEZOELECTRIC SHELL COMPOSITES

Mathematical models of the flexible beams and plates (rectangular and square plates) are derived from a generic theory of nonlinear thin anisotropic piezothermoelastic shells. A generic anisotropic deep piezothermoelastic shell is defined in a curvilinear coordinate system, and it is exposed to mechanical, electric, and thermal excitations. Figure 1 illustrates the original generic shell and its derivative geometries including plates, beams, and other shell, non-shell geometries. It is assumed that the shell is subjected to a large deformation resulting in a geometrical nonlinearity. However, material properties are assumed constant, and the stress and strain relations are linear. The generic theory is derived based on a generic anisotropic piezothermoelastic thin shell. Simplification of the generic theory to other piezoelectric materials, e.g., mm2 (polyvinylidene fluoride), mm6 (piezoceramics), etc., or piezoelectric continua, e.g., spherical shells, cylindrical shells, plates, beams, etc., can be achieved when appropriate material or geometrical parameters are defined (Tzou, 1993). A generic infinitesimal distance  $ds$  in a shell can be defined by a *fundamental form*:

$$(ds)^2 = \sum_{i=1}^3 (f_{ii})^2 (d\alpha_i)^2, \quad (17)$$

where  $f_{ii}(\alpha_1, \alpha_2, \alpha_3) = A_i(1 + \frac{\alpha_3}{R_i})$  ( $i=1,2$ ),  $f_{33}(\alpha_1, \alpha_2, \alpha_3) = 1$ ;  $\alpha_3$  is a finite distance measured from the reference surface;  $A_1$  and  $A_2$  are the Lamé parameters; and  $R_1$  and  $R_2$  are the radii of curvature of the  $\alpha_1$  and  $\alpha_2$  axes on the surface defined by  $\alpha_3 = 0$ . For beams and rectangular plates,  $A_1 = A_2 = 1$  and  $R_1 = R_2 = \infty$ . Thus,  $(ds)^2 = \sum_{i=1}^3 (1)^2 (d\alpha_i)^2$ . For convenience, the neutral surface is taken as the reference surface, which is defined by the  $\alpha_1$  and  $\alpha_2$  axes.

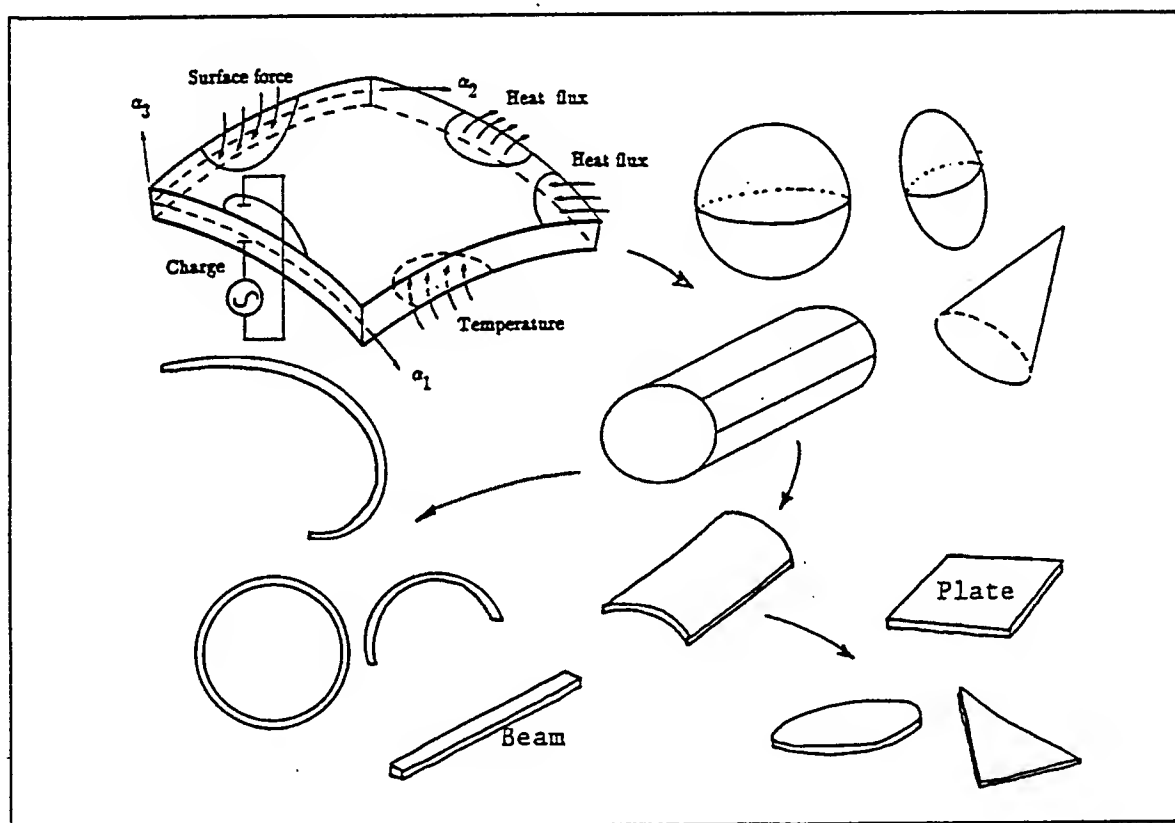


Fig.1 A nonlinear piezothermoelastic shell and its derivative geometries.

Hamilton's principle is used in deriving the system equations and boundary conditions of the piezothermoelastic shell continuum. Simplifying the shell equations gives the nonlinear equations of flexible beams and plates. Hamilton's principle assumes that the energy variations over an arbitrary period of time are zero. Substituting all energies into Hamilton's equation and carrying out all derivations, one can derive the nonlinear

piezothermoelastic equations and boundary conditions of the nonlinear piezothermoelastic shell (Tzou and Bao, 1996).

$$\frac{\partial(N_{11}A_2)}{\partial\alpha_1} - N_{22}\frac{\partial A_2}{\partial\alpha_1} + \frac{\partial(N_{12}A_1)}{\partial\alpha_2} + N_{12}\frac{\partial A_1}{\partial\alpha_2} + Q_{13}\frac{A_1A_2}{R_1} = A_1A_2\rho h\ddot{u}_1, \quad (18)$$

$$\frac{\partial(N_{12}A_2)}{\partial\alpha_1} + N_{12}\frac{\partial A_2}{\partial\alpha_1} + \frac{\partial(N_{22}A_1)}{\partial\alpha_2} - N_{11}\frac{\partial A_1}{\partial\alpha_2} + Q_{23}\frac{A_1A_2}{R_2} = A_1A_2\rho h\ddot{u}_2, \quad (19)$$

$$\begin{aligned} & \frac{\partial(Q_{13}A_2)}{\partial\alpha_1} + \frac{\partial(Q_{23}A_1)}{\partial\alpha_2} + A_1A_2 \sum_{k=1}^n \bar{T}_{3k} \left| \frac{\alpha_{3k}}{\alpha_{3k-1}} - A_1A_2 \left[ \frac{N_{11}}{R_1} + \frac{N_{22}}{R_2} \right] \right. \\ & + \left\{ \left[ \frac{\partial}{\partial\alpha_1}(N_{11}\frac{A_2}{A_1}) + \frac{\partial}{\partial\alpha_2}(N_{12}) \right] \frac{\partial u_3}{\partial\alpha_1} + \left[ \frac{\partial}{\partial\alpha_2}(N_{22}\frac{A_1}{A_2}) + \frac{\partial}{\partial\alpha_1}(N_{12}) \right] \frac{\partial u_3}{\partial\alpha_2} \right. \\ & \left. \left. + 2N_{12} \frac{\partial^2 u_3}{\partial\alpha_1 \partial\alpha_2} + N_{11} \frac{A_2}{A_1} \frac{\partial^2 u_3}{\partial\alpha_1^2} + N_{22} \frac{A_1}{A_2} \frac{\partial^2 u_3}{\partial\alpha_2^2} \right\} = A_1A_2\rho h\ddot{u}_3 \end{aligned} \quad (20)$$

where  $Q_{13}$  and  $Q_{23}$  are defined by

$$Q_{13} = \frac{1}{A_1A_2} \left[ \frac{\partial(M_{11}A_2)}{\partial\alpha_1} - M_{22}\frac{\partial A_2}{\partial\alpha_1} + \frac{\partial(M_{12}A_1)}{\partial\alpha_2} + M_{12}\frac{\partial A_1}{\partial\alpha_2} \right], \quad (21)$$

$$Q_{23} = \frac{1}{A_1A_2} \left[ \frac{\partial(M_{12}A_2)}{\partial\alpha_1} + M_{12}\frac{\partial A_2}{\partial\alpha_1} + \frac{\partial(M_{22}A_1)}{\partial\alpha_2} - M_{11}\frac{\partial A_1}{\partial\alpha_2} \right], \quad (22)$$

where  $\rho = \frac{1}{h} \sum_{k=1}^n \rho_k h_k$  is defined as a weighted average density for the multi-layered shell.

It is observed that the **nonlinear influence** on the transverse equation  $u_3$  is very prominent. (All terms inside the brace are contributed by the nonlinear effects of the von Karman type geometric nonlinearity.) Note that the thermo-electromechanical equations look like a standard shell equations. However, the force and moment expressions defined by mechanical, thermal, and electric effects are much more complicated than the conventional elastic expressions. Substituting the expressions of  $N_{11}$ ,  $N_{22}$ ,  $N_{12}$ ,  $M_{11}$ ,  $M_{22}$ ,  $M_{12}$  into the above equations leads to the thermo-electromechanical equations defined in the reference displacements  $u_1$ ,  $u_2$ ,  $u_3$ . The transverse shear deformation and rotatory inertia effects are not considered. The electric terms, forces and moments, can be used in controlling the mechanical and/or temperature induced excitations (Tzou and Ye, 1994). For nonlinear flexible beams and plates (rectangular and square plates),  $A_1 = A_2 = 1$  and  $R_1 = R_2 = \infty$ . Substituting the Lamé parameters and radii into the shell thermo-electromechanical equations and simplifying, one can derive the governing equations for nonlinear

piezoelectric laminated beams and plates. Accordingly, thermo-electromechanical couplings and control of static/dynamic nonlinearities can be investigated. Detailed derivations are presented next.

## NONLINEAR PIEZOELECTRIC COMPOSITE PLATES

For piezoelectric laminated composite rectangular plates, Figure 2, the coordinate system, radii of curvatures, and Lamé parameters are defined as follows:  $\alpha_1 = x$ ,  $\alpha_2 = y$ ,  $\alpha_3 = z$ ,  $R_1 = \infty$ ,  $R_2 = \infty$ ,  $A_1 = 1$ , and  $A_2 = 1$ .

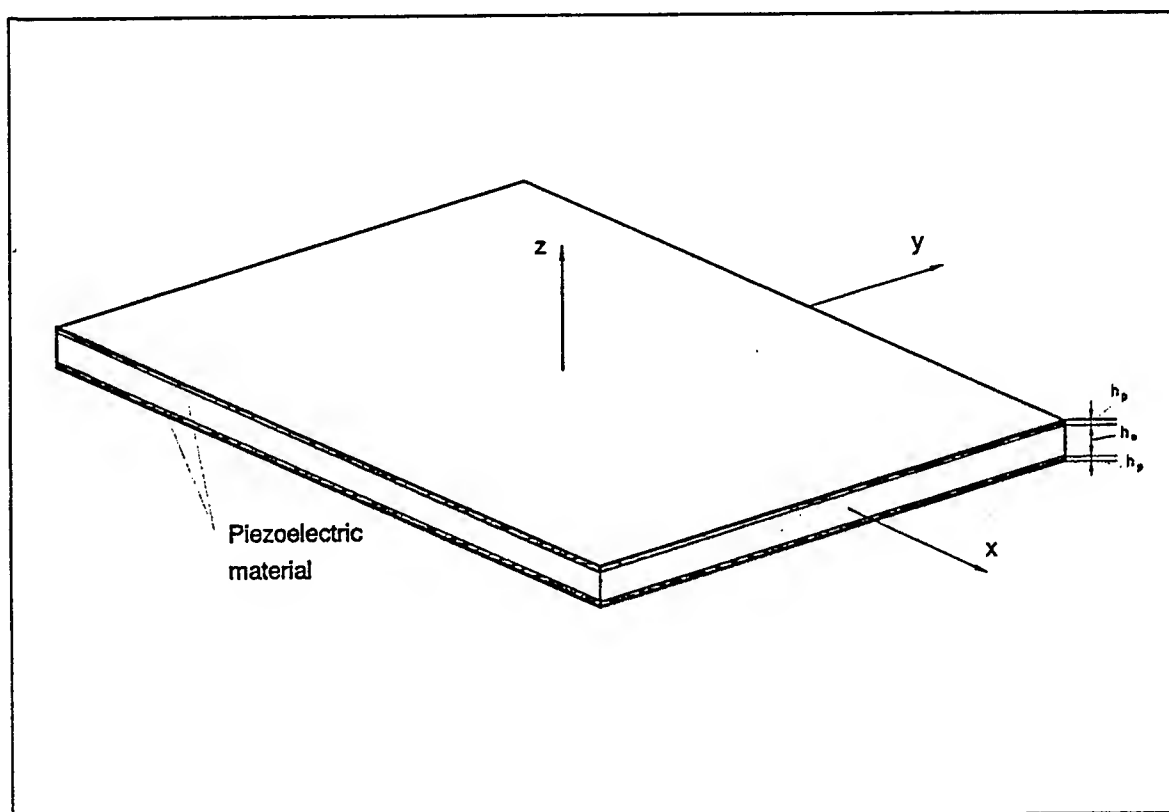


Fig.2 Nonlinear composite plate.

It is assumed that an elastic plate, with dimensions  $2a \times 2b \times h$ , is sandwiched between two piezoelectric layers and the composite plate experiences a large deformation. The resultant membrane forces and the bending moments are

$$\begin{bmatrix} N_{xx} \\ N_{yy} \\ N_{xy} \end{bmatrix} = \begin{bmatrix} N_{xx}^m \\ N_{yy}^m \\ N_{xy}^m \end{bmatrix} - \begin{bmatrix} N_{xx}^e \\ N_{yy}^e \\ N_{xy}^e \end{bmatrix} - \begin{bmatrix} N_{xx}^\theta \\ N_{yy}^\theta \\ N_{xy}^\theta \end{bmatrix}; \quad \begin{bmatrix} M_{xx} \\ M_{yy} \\ M_{xy} \end{bmatrix} = \begin{bmatrix} M_{xx}^m \\ M_{yy}^m \\ M_{xy}^m \end{bmatrix} - \begin{bmatrix} M_{xx}^e \\ M_{yy}^e \\ M_{xy}^e \end{bmatrix} - \begin{bmatrix} M_{xx}^\theta \\ M_{yy}^\theta \\ M_{xy}^\theta \end{bmatrix}, \quad (23)$$

where the superscript "m" denotes the mechanically induced components, "e" the electrically induced components and " $\theta$ " the thermally induced component. The mechanical membrane forces and moments are

$$\begin{bmatrix} N_{xx}^m \\ N_{yy}^m \\ N_{xy}^m \end{bmatrix} = K \begin{bmatrix} s_x^* + \mu s_y^* \\ s_y^* + \mu s_x^* \\ s_{xy}^*(1-\mu)/2 \end{bmatrix}, \quad (24)$$

where  $K = \frac{Yh}{1-\mu^2}$ ;  $h=2d$  is the thickness of elastic plate;

$\mu$  is Poisson's ratio, and  $Y$  is the Young's modulus;

$$\begin{bmatrix} M_{xx}^m \\ M_{yy}^m \\ M_{xy}^m \end{bmatrix} = D \begin{bmatrix} \kappa_x + \mu\kappa_y \\ \kappa_y + \mu\kappa_x \\ \kappa_{xy}(1-\mu)/2 \end{bmatrix}, \quad \text{where } D = \frac{Yh^3}{12(1-\mu^2)}; \quad \text{and}$$

$$\begin{bmatrix} N_{xx}^e \\ N_{yy}^e \\ N_{xy}^e \end{bmatrix} = -e_{31}(\phi_{31} + \phi_{33}) \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \quad \text{where } \phi_{3j} \text{ is the total voltage across the } j\text{-th layer};$$

$$\begin{bmatrix} M_{xx}^e \\ M_{yy}^e \\ M_{xy}^e \end{bmatrix} = \frac{e_{31}(h+t)}{2}(\phi_{31} - \phi_{33}) \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \quad \text{where } t \text{ is the thickness of piezoelectric layers};$$

$$\begin{bmatrix} N_{xx}^\theta \\ N_{yy}^\theta \\ N_{xy}^\theta \end{bmatrix} = \int_{\alpha_3} \begin{bmatrix} \lambda_x \\ \lambda_y \\ 0 \end{bmatrix} \theta d\alpha_3 = \int_{-d}^{+d} \begin{bmatrix} \lambda_x \\ \lambda_y \\ 0 \end{bmatrix} \theta d\alpha_3;$$

$$\begin{bmatrix} M_{xx}^\theta \\ M_{yy}^\theta \\ M_{xy}^\theta \end{bmatrix} = \int_{\alpha_3} \begin{bmatrix} \lambda_x \\ \lambda_y \\ 0 \end{bmatrix} \theta \alpha_3 d\alpha_3 = \int_{-d}^{+d} \begin{bmatrix} \lambda_x \\ \lambda_y \\ 0 \end{bmatrix} \theta \alpha_3 d\alpha_3.$$

The relations of strain-displacement are

$$\begin{bmatrix} S_x \\ S_y \\ S_{xy} \end{bmatrix} = \begin{bmatrix} s_x^* \\ s_y^* \\ s_{xy}^* \end{bmatrix} + z \begin{bmatrix} \kappa_x \\ \kappa_y \\ \kappa_{xy} \end{bmatrix}, \quad (25)$$

where the membrane and bending strains are

$$\begin{aligned}
s^{\circ}_x &= \frac{\partial u_x}{\partial x} + \frac{1}{2} \left( \frac{\partial u_z}{\partial x} \right)^2; \quad \kappa_x = \frac{\partial \beta_x}{\partial x} = -\frac{\partial^2 u_z}{\partial x^2}; \\
s^{\circ}_y &= \frac{\partial u_y}{\partial y} + \frac{1}{2} \left( \frac{\partial u_z}{\partial y} \right)^2; \quad \kappa_y = \frac{\partial \beta_y}{\partial y} = -\frac{\partial^2 u_z}{\partial y^2}; \\
s^{\circ}_{xy} &= \frac{\partial u_x}{\partial y} + \frac{\partial u_y}{\partial x} + \frac{\partial u_z}{\partial x} \frac{\partial u_z}{\partial y}; \quad \kappa_{xy} = \frac{\partial \beta_x}{\partial y} + \frac{\partial \beta_y}{\partial x} = -2 \frac{\partial^2 u_z}{\partial x \partial y}.
\end{aligned} \tag{26}$$

Note that the quadratic terms in the membrane strains are the von Karman-type nonlinear terms. The inplane inertia effects of thin plates are usually neglected and the in-plane mechanical forces are assumed zero, i.e.,  $q_x = q_y = 0$ . Then, the nonlinear plate equations are derived as

$$\frac{\partial N_{xx}}{\partial x} + \frac{\partial N_{xy}}{\partial y} = 0; \quad \frac{\partial N_{xy}}{\partial x} + \frac{\partial N_{yy}}{\partial y} = 0; \tag{27a,b}$$

$$\frac{\partial^2 M_{xx}}{\partial x^2} + 2 \frac{\partial^2 M_{xy}}{\partial x \partial y} + \frac{\partial^2 M_{yy}}{\partial y^2} + N_{xx} \frac{\partial^2 u_z}{\partial x^2} + 2 N_{xy} \frac{\partial^2 u_z}{\partial x \partial y} + N_{yy} \frac{\partial^2 u_z}{\partial y^2} + q_z = \rho h \ddot{u}_z. \tag{28}$$

The static version of these equations are those of the von Kármán plate theory. These simplified static equations are identical to those in (Chai, 1980).

Substituting all force and moment expressions (mechanical, thermal, and electric components) into the plate equations yields three piezothermoelastic equations in terms of the displacements  $u_x$ ,  $u_y$  and  $u_z$ .

$$\frac{\partial N_{xx}^m}{\partial x} + \frac{\partial N_{xy}^m}{\partial y} = \left( \frac{\partial N_{xx}^e}{\partial x} + \frac{\partial N_{xx}^{\theta}}{\partial x} \right) + \left( \frac{\partial N_{xy}^e}{\partial y} + \frac{\partial N_{xy}^{\theta}}{\partial y} \right); \tag{29}$$

$$\frac{\partial N_{xy}^m}{\partial x} + \frac{\partial N_{yy}^m}{\partial y} = \left( \frac{\partial N_{xy}^e}{\partial x} + \frac{\partial N_{xy}^{\theta}}{\partial x} \right) + \left( \frac{\partial N_{yy}^e}{\partial y} + \frac{\partial N_{yy}^{\theta}}{\partial y} \right); \tag{30}$$

$$\begin{aligned}
&\frac{\partial^2 M_{xx}^m}{\partial x^2} + 2 \frac{\partial^2 M_{xy}^m}{\partial x \partial y} + \frac{\partial^2 M_{yy}^m}{\partial y^2} - \rho h \ddot{u}_z = -q_z \\
&- (N_{xx}^m \frac{\partial^2 u_z}{\partial x^2} + 2 N_{xy}^m \frac{\partial^2 u_z}{\partial x \partial y} + N_{yy}^m \frac{\partial^2 u_z}{\partial y^2}) \\
&+ [(N_{xx}^e + N_{xx}^{\theta}) \frac{\partial^2 u_z}{\partial x^2} + 2(N_{xy}^e + N_{xy}^{\theta}) \frac{\partial^2 u_z}{\partial x \partial y} + (N_{yy}^e + N_{yy}^{\theta}) \frac{\partial^2 u_z}{\partial y^2}] \\
&+ [(\frac{\partial^2 M_{xx}^e}{\partial x^2} + \frac{\partial^2 M_{xx}^{\theta}}{\partial x^2}) + (2 \frac{\partial^2 M_{xy}^e}{\partial x \partial y} + 2 \frac{\partial^2 M_{xy}^{\theta}}{\partial x \partial y}) + (\frac{\partial^2 M_{yy}^e}{\partial y^2} + \frac{\partial^2 M_{yy}^{\theta}}{\partial y^2})].
\end{aligned} \tag{31}$$



Furthermore, substituting all force and moment expressions, one can derive the displacement equations as

$$\begin{aligned} \frac{\partial^2 u_x}{\partial x^2} + \eta_1 \frac{\partial^2 u_x}{\partial y^2} + \eta_2 \frac{\partial^2 u_y}{\partial x \partial y} = & -\frac{\partial u_z}{\partial x} \left( \frac{\partial^2 u_z}{\partial x^2} + \eta_1 \frac{\partial^2 u_z}{\partial y^2} \right) \\ & - \eta_2 \frac{\partial u_z}{\partial y} \frac{\partial^2 u_z}{\partial x \partial y} + \frac{1}{K} \left[ \left( \frac{\partial N_{xx}^e}{\partial x} + \frac{\partial N_{xx}^\theta}{\partial x} \right) + \left( \frac{\partial N_{xy}^e}{\partial y} + \frac{\partial N_{xy}^\theta}{\partial y} \right) \right]; \end{aligned} \quad (32)$$

$$\begin{aligned} \eta_2 \frac{\partial^2 u_x}{\partial x \partial y} + \eta_1 \frac{\partial^2 u_y}{\partial x^2} + \frac{\partial^2 u_y}{\partial y^2} = & -\frac{\partial u_z}{\partial y} \left( \frac{\partial^2 u_z}{\partial y^2} + \eta_1 \frac{\partial^2 u_z}{\partial x^2} \right) \\ & - \eta_2 \frac{\partial u_z}{\partial x} \frac{\partial^2 u_z}{\partial x \partial y} + \frac{1}{K} \left[ \left( \frac{\partial N_{xy}^e}{\partial x} + \frac{\partial N_{xy}^\theta}{\partial x} \right) + \left( \frac{\partial N_{yy}^e}{\partial y} + \frac{\partial N_{yy}^\theta}{\partial y} \right) \right]; \end{aligned} \quad (33)$$

$$\begin{aligned} D \nabla^2 \nabla^2 u_z + \rho h \ddot{u}_z = & q_z + K \left[ (1-\mu) \left( \frac{\partial u_x}{\partial y} + \frac{\partial u_y}{\partial x} + \frac{\partial u_z}{\partial x} \frac{\partial u_z}{\partial y} \frac{\partial^2 u_z}{\partial x \partial y} \right. \right. \\ & + \left. \left( \frac{\partial u_x}{\partial x} + \frac{1}{2} \left( \frac{\partial u_z}{\partial x} \right)^2 \right) \left( \frac{\partial^2 u_z}{\partial x^2} + \mu \frac{\partial^2 u_z}{\partial y^2} \right) + \left( \frac{\partial u_y}{\partial y} + \frac{1}{2} \left( \frac{\partial u_z}{\partial y} \right)^2 \right) \left( \frac{\partial^2 u_z}{\partial y^2} + \mu \frac{\partial^2 u_z}{\partial x^2} \right) \right] \\ & - \left[ (N_{xx}^e + N_{xx}^\theta) \frac{\partial^2 u_z}{\partial x^2} + 2(N_{xy}^e + N_{xy}^\theta) \frac{\partial^2 u_z}{\partial x \partial y} + (N_{yy}^e + N_{yy}^\theta) \frac{\partial^2 u_z}{\partial y^2} \right] \\ & - \left[ \left( \frac{\partial^2 M_{xx}^e}{\partial x^2} + \frac{\partial^2 M_{xx}^\theta}{\partial x^2} \right) + \left( 2 \frac{\partial^2 M_{xy}^e}{\partial x \partial y} + 2 \frac{\partial^2 M_{xy}^\theta}{\partial x \partial y} \right) + \left( \frac{\partial^2 M_{yy}^e}{\partial y^2} + \frac{\partial^2 M_{yy}^\theta}{\partial y^2} \right) \right]. \end{aligned} \quad (34)$$

where  $\eta_1 = \frac{1-\mu}{2}$ ,  $\eta_2 = \frac{1+\mu}{2}$ ,  $\nabla^2$  is the Laplacian operator, and  $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ .

For a uniformly distributed piezoelectric layer with constant thickness, the electric potentials on the piezo-layers are independent of the coordinates  $x$  and  $y$ . It is also assumed that the temperature rise is uniform with respect to the  $x$  and  $y$  coordinates,  $\theta(x,y,z) \equiv \theta(z)$ . Occasionally, introducing a generic forcing function and representing the existing force terms by the generic forcing function in certain boundary-value problems would significantly simplify the system equations and also alleviate the complexity of solution procedures. Thus, a generic forcing function  $\hat{F}(x,y)$  is introduced and the forces are defined by

$$N_{xx}^m = \frac{\partial^2 \hat{F}}{\partial y^2}, \quad N_{yy}^m = \frac{\partial^2 \hat{F}}{\partial x^2}, \quad \text{and} \quad N_{xy}^m = -\frac{\partial^2 \hat{F}}{\partial x \partial y}. \quad (35)$$

Note that the two in-plane equations are exactly satisfied and the transverse equation is expressed in terms of the transverse deflection  $u_3$  and the forcing function  $\hat{F}$ .

$$D \nabla^2 \nabla^2 u_z + \rho h \ddot{u}_z = q_z + \hat{N}(\hat{F}, u_z) - [(N_{xx}^e + N_{xx}^\theta) \frac{\partial^2 u_z}{\partial x^2} + 2(N_{xy}^e + N_{xy}^\theta) \frac{\partial^2 u_z}{\partial x \partial y} + (N_{yy}^e + N_{yy}^\theta) \frac{\partial^2 u_z}{\partial y^2}] , \quad (36)$$

where  $\hat{N}(\hat{F}, u_z)$  is the nonlinear operator defined by

$$\hat{N}(\hat{F}, u_z) = \frac{\partial^2 \hat{F}}{\partial y^2} \frac{\partial^2 u_z}{\partial x^2} + \frac{\partial^2 \hat{F}}{\partial x^2} \frac{\partial^2 u_z}{\partial y^2} - 2 \frac{\partial^2 \hat{F}}{\partial x \partial y} \frac{\partial^2 u_z}{\partial x \partial y} . \quad (37)$$

The second equation is the compatibility equation of the plate. Applying the membrane strain equations leads to the compatibility condition:

$$\frac{\partial^2 s^{\circ}_x}{\partial y^2} + \frac{\partial^2 s^{\circ}_y}{\partial x^2} - \frac{\partial^2 s^{\circ}_{xy}}{\partial x \partial y} = \left( \frac{\partial^2 u_z}{\partial x \partial y} \right)^2 - \frac{\partial^2 u_z}{\partial x^2} \frac{\partial^2 u_z}{\partial y^2} . \quad (38)$$

Since the forces are functions of membrane strains  $\begin{bmatrix} N_{xx}^m \\ N_{yy}^m \\ N_{xy}^m \end{bmatrix} = K \begin{bmatrix} 1 & \mu & 0 \\ \mu & 1 & 0 \\ 0 & 0 & (1-\mu)/2 \end{bmatrix} \begin{bmatrix} s^{\circ}_x \\ s^{\circ}_y \\ s^{\circ}_{xy} \end{bmatrix}$ , the strains can be expressed as functions of forces

$$\begin{bmatrix} s^{\circ}_x \\ s^{\circ}_y \\ s^{\circ}_{xy} \end{bmatrix} = \frac{1}{K} \begin{bmatrix} 1 & \mu & 0 \\ \mu & 1 & 0 \\ 0 & 0 & (1-\mu)/2 \end{bmatrix}^{-1} \begin{bmatrix} N_{xx}^m \\ N_{yy}^m \\ N_{xy}^m \end{bmatrix} = \frac{1}{K(1-\mu^2)} \begin{bmatrix} 1 & -\mu & 0 \\ -\mu & 1 & 0 \\ 0 & 0 & 2(1+\mu) \end{bmatrix} \begin{bmatrix} N_{xx}^m \\ N_{yy}^m \\ N_{xy}^m \end{bmatrix} . \quad (39)$$

Referring to the forces represented by the generic forcing function, i.e.,  $N_{xx}^m = \frac{\partial^2 \hat{F}}{\partial y^2}, \dots$ , one can also define the strains as functions of the forcing function.

$$s^{\circ}_x = \frac{1}{Y_h} (N_{xx}^m - \mu N_{yy}^m) = \frac{1}{Y_h} \left( \frac{\partial^2 \hat{F}}{\partial y^2} - \mu \frac{\partial^2 \hat{F}}{\partial x^2} \right) ; \quad (40)$$

$$s^{\circ}_y = \frac{1}{Y_h} (N_{yy}^m - \mu N_{xx}^m) = \frac{1}{Y_h} \left( \frac{\partial^2 \hat{F}}{\partial x^2} - \mu \frac{\partial^2 \hat{F}}{\partial y^2} \right) ; \quad (41)$$

$$s^{\circ}_{xy} = \frac{2(1+\mu)}{Y_h} N_{xy}^m = -\frac{2(1+\mu)}{Y_h} \frac{\partial^2 \hat{F}}{\partial x \partial y} . \quad (42)$$

The second equation then can be obtained by substituting these equations into the condition of compatibility:

$$\nabla^2 \nabla^2 \hat{F}(x,y) + \frac{Yh}{2} \hat{N}(u_z, u_z) = 0. \quad (43)$$

Since the closed-form solutions to these nonlinear equations are usually not available, the approximation methods are used to solve these equations. The approximation techniques include the (generalized) double Fourier method, Rayleigh-Ritz method, Galerkin's method, perturbation technique and finite difference technique, etc. Simulation results based on the series solutions are presented later.

Four boundary conditions are usually required for each edge of the plate. Electric boundary conditions are also allowed and from which boundary control can be implemented. Generic boundary conditions are

$$\begin{array}{ll} u_n = u_n^* & \text{or} \quad N_{nn} = N_{nn}^*; \\ u_s = u_s^* & \text{or} \quad N_{ns} = N_{ns}^*; \\ u_z = u_z^* & \text{or} \quad \frac{\partial M_{ns}}{\partial s} + Q_n = \frac{\partial M_{ns}^*}{\partial s} + Q_n^*; \\ \frac{\partial u_z}{\partial n} = \frac{\partial u_z^*}{\partial n} & \text{or} \quad M_{nn} = M_{nn}^*, \end{array} \quad (44)$$

in which the subscripts n and s denote the directions outward normal and tangential to the boundary, respectively, and the starred quantities designate the prescribed value. Typical mechanical boundary conditions for the plate are:

- a) Rigidly clamped edge:  $u_z = \frac{\partial u_z}{\partial n} = u_n = u_s = 0$ ; (45a-e)
- b) Loosely clamped edge:  $u_z = \frac{\partial u_z}{\partial n} = N_{nn} = N_{ns} = 0$ ;
- c) Simply supported edge (movable in the plane of the plate)  
 $u_z = M_{nn} = N_{nn} = N_{ns} = 0$ ;
- d) Hinged edge (simply supported edge immovable in the plane of the plate):  
 $u_z = M_{nn} = u_n = u_s = 0$ ;
- e) Free edge:  $N_{nn} = N_{ns} = M_{nn} = Q_n + \frac{\partial M_{ns}}{\partial s} = 0$ .

## NONLINEAR PIEZOELECTRIC COMPOSITE BEAMS

A piezothermoelastic laminated beam is shown in Figure 3 in which two piezoelectric layers, dimensions  $L \times b \times h_p$ , are perfectly bounded on the top and the bottom surfaces of a steel beam, dimensions  $L \times b \times h_e$ . Thus, the total laminated beam thickness is  $h = h_e + 2h_p$ . It is assumed that the laminated beam undergoes a von Karman type geometrical nonlinearity and temperature and electric inputs.

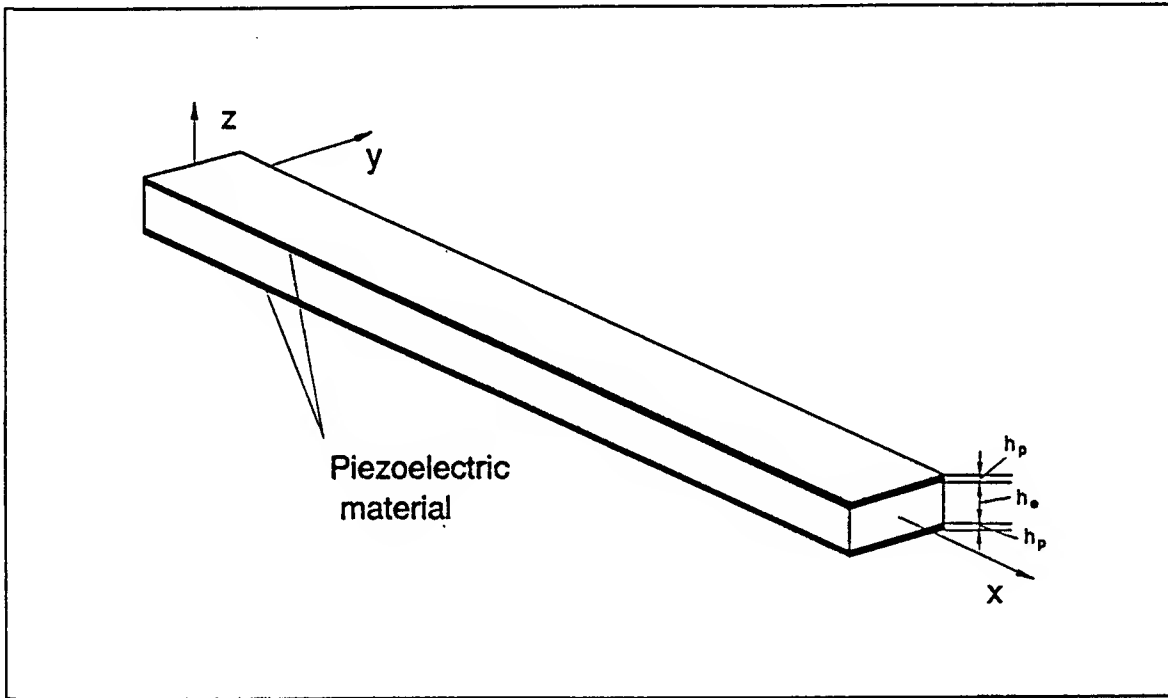


Fig.3 A piezothermoelastic laminated beam.

Simplifying the governing equations of the nonlinear piezothermoelastic shell laminate (Tzou and Bao, 1995), one obtains the nonlinear piezothermoelastic beam equations in the longitudinal (x) direction and the transverse (z) direction, respectively.

$$\frac{\partial N_{xx}}{\partial x} + q_x = \rho h \ddot{u}_x, \quad (46)$$

$$\frac{\partial^2 M_{xx}}{\partial x^2} + \frac{\partial N_{xx}}{\partial x} \frac{\partial u_z}{\partial x} + N_{xx} \frac{\partial^2 u_z}{\partial x^2} + q_z = \rho h \ddot{u}_z, \quad (47)$$

where the mass per unit length  $\rho h = \Sigma \int \rho_k dz = 2\rho_p h_p + \rho_e h_e$ ;  $\rho_p$  and  $\rho_e$  are the densities of the piezoelectric layer and the elastic steel layer, respectively.  $N_{xx}$  and  $M_{xx}$  are the membrane force and bending moment per unit width. Since the beam width  $b$  is constant, one can define  $N_x = bN_{xx}$  and  $M_x = bM_{xx}$ . Note that the forces and moments include all elastic, electric, and temperature effects.

$$N_x = (Y\tilde{A} + 2Y_p\tilde{A}_p + 2\tilde{A}_p\frac{e_{31}^2}{\epsilon_{33}})s_{xx}^* + e_{31}b(\phi_{33}^c + \phi_{31}^c) + \left[ \left( \frac{e_{31}p_3}{\epsilon_{33}} - \lambda_p \right) \left[ \int_{-(h_e/2+h_p)}^{-h_e/2} \theta b dz + \int_{h_e/2}^{h_e/2+h_p} \theta b dz \right] - \lambda \int_{-h_e/2}^{h_e/2} \theta b dz \right], \quad (48a)$$

$$M_x = (YI + 2Y_pI_p + 2I_p\frac{e_{31}^2}{\epsilon_{33}})\kappa_{xx} + e_{31}bra(\phi_{33}^c - \phi_{31}^c) + \left[ \left( \frac{e_{31}p_3}{\epsilon_{33}} - \lambda_p \right) \left[ \int_{-(h_e/2+h_p)}^{-h_e/2} \theta z b dz + \int_{h_e/2}^{h_e/2+h_p} \theta z b dz \right] - \lambda \int_{-h_e/2}^{h_e/2} \theta z b dz \right], \quad (48b)$$

where  $\tilde{A} = bh_e$ ,  $\tilde{A}_p = bh_p$  are the cross section areas of the elastic steel layer and the piezoelectric layers, respectively;  $Y$  and  $Y_p$  are Young's moduli of the steel and the piezoelectric material;  $I = bh_e^3/12$ ,  $I_p = bh_p^3/12 + bh_p(h_e+h_p)^2/2$  are the area moments of the steel layer and the piezoelectric layer, respectively;  $\phi_{33}^c$  and  $\phi_{31}^c$  are the control voltages applied to the top and the bottom piezoelectric layers;  $\theta$  is the temperature variation;  $e_{31}$ ,  $\epsilon_{33}$ ,  $p_3$  and  $\lambda_p$  are the piezoelectric stress coefficient, the dielectric coefficient, the pyroelectric coefficient, and the stress-temperature coefficient for the piezoelectric material, respectively;  $\lambda$  is the steel stress-temperature coefficient;  $s_{xx}^*$  and  $\kappa_{xx}$  are the membrane strain and bending strain; and  $ra = (h_e+h_p)/2$  is the actuator moment arm. Also, one can define  $p_x = bq_x$ ,  $p_z = bq_z$  (mechanical excitations per unit length), and  $\bar{m} = \rho bh = 2\rho_p\tilde{A}_p + \rho\tilde{A}$  (mass per unit beam length). Then, Eqs.(46,47b) can be rewritten as

$$\frac{\partial N_x}{\partial x} + p_x = \bar{m}\ddot{u}_x, \quad (49)$$

$$\frac{\partial^2 M_x}{\partial x^2} + \frac{\partial N_x}{\partial x} \frac{\partial u_z}{\partial x} + N_x \frac{\partial^2 u_z}{\partial x^2} + p_z = \bar{m}\ddot{u}_z. \quad (50)$$

Using Eqs.(48a&b), one can write the axial force and bending moment in a compact form:

$$N_x = \tilde{K} s_{xx}^\circ + N_x^c + N_x^t, \quad (51a)$$

$$M_x = \tilde{D} \kappa_{xx} + M_x^c + M_x^t, \quad (51b)$$

where  $\tilde{K}$  is the membrane stiffness  $\tilde{K} = (Y\tilde{A} + 2Y_p\tilde{A}_p + 2\tilde{A}_p\frac{e_{31}^2}{\epsilon_{33}})$ ;  $\tilde{D}$  is the bending stiffness  $\tilde{D} = (YI + 2Y_pI_p + 2I_p\frac{e_{31}^2}{\epsilon_{33}})$ ;  $s_{xx}^\circ$  is the membrane strain and  $\kappa_{xx}$  is the bending strain with the von Karman type nonlinearity:

$$s_{xx}^\circ = \frac{\partial u_x}{\partial x} + \frac{1}{2} \left( \frac{\partial u_z}{\partial x} \right)^2, \quad \kappa_{xx} = \frac{\partial \beta_x}{\partial x} = -\frac{\partial^2 u_z}{\partial x^2}. \quad (52a\&b)$$

$N_x^t$  is the axial force induced by the temperature rise;  $M_x^t$  is the moment due to the temperature rise;  $N_x^c$  is the axial force induced by the control potential; and  $M_x^c$  is the control moment induced by the control potential.

$$N_x^c = e_{31}b(\phi_{33}^c + \phi_{31}^c), \quad (53a)$$

$$M_x^c = e_{31}bra(\phi_{33}^c - \phi_{31}^c), \quad (53b)$$

$$N_x^t = \left[ \left( \frac{e_{31}p_3}{\epsilon_{33}} - \lambda_p \right) \left[ \int_{-(h_e/2+h_p)}^{-h_e/2} \theta_b dz + \int_{h_e/2}^{h_e/2+h_p} \theta_b dz \right] - \lambda \int_{-h_e/2}^{h_e/2} \theta_b dz \right], \quad (53c)$$

$$M_x^t = \left[ \left( \frac{e_{31}p_3}{\epsilon_{33}} - \lambda_p \right) \left[ \int_{-(h_e/2+h_p)}^{-h_e/2} \theta_z b dz + \int_{h_e/2}^{h_e/2+h_p} \theta_z b dz \right] - \lambda \int_{-h_e/2}^{h_e/2} \theta_z b dz \right]. \quad (53d)$$

Boundary conditions at the two ends of the laminated beam,  $x = 0$  and  $x = L$ , are

$$N_x = N_x^* \quad \text{or} \quad u_x = u_x^*, \quad (54a)$$

$$M_x = M_x^* \quad \text{or} \quad \beta_x = \beta_x^*, \quad (54b)$$

$$\frac{\partial M_x}{\partial x} + N_x \frac{\partial u_z}{\partial x} = Q_z^* \quad \text{or} \quad u_z = u_z^*, \quad (54c)$$

where the quantities with the asterisk "\*" are prescribed values on the boundary. Note that usually either force boundary conditions or displacement boundary conditions are

selected for a given physical boundary condition.

It is assumed that the mechanical excitations in the longitudinal and transverse directions are zero, i.e.,  $p_z = p_x = 0$ . Substituting the axial force and bending moment into Eqs.(49,50), one can write

$$\tilde{K} \frac{\partial}{\partial x} \left[ \frac{\partial u_x}{\partial x} + \frac{1}{2} \left( \frac{\partial u_z}{\partial x} \right)^2 \right] + e_{31} b \frac{\partial}{\partial x} (\phi_{33}^c + \phi_{31}^c) + \left[ \left( \frac{e_{31} p_3}{\epsilon_{33}} - \lambda_p \right) \cdot \left( \int_{-(h_e/2+h_p)}^{-h_e/2} \frac{\partial \theta}{\partial x} b dz + \int_{h_e/2}^{h_e/2+h_p} \frac{\partial \theta}{\partial x} b dz \right) - \lambda \int_{-h_e/2}^{h_e/2} \frac{\partial \theta}{\partial x} b dz \right] = \bar{m} \ddot{u}_x ; \quad (55a)$$

$$\begin{aligned} & \left\{ -\tilde{D} \frac{\partial^4 u_z}{\partial x^4} + e_{31} b r^a \frac{\partial^2}{\partial x^2} (\phi_{33}^c - \phi_{31}^c) + \left[ \left( \frac{e_{31} p_3}{\epsilon_{33}} - \lambda_p \right) \cdot \left( \int_{-(h_e/2+h_p)}^{-h_e/2} \frac{\partial^2 \theta}{\partial x^2} z b dz + \int_{h_e/2}^{h_e/2+h_p} \frac{\partial^2 \theta}{\partial x^2} z b dz \right) - \lambda \int_{-h_e/2}^{h_e/2} \frac{\partial^2 \theta}{\partial x^2} z b dz \right] \right\} \\ & + \left\{ \tilde{K} \frac{\partial}{\partial x} \left[ \frac{\partial u_x}{\partial x} + \frac{1}{2} \left( \frac{\partial u_z}{\partial x} \right)^2 \right] + e_{31} b \frac{\partial}{\partial x} (\phi_{33}^c + \phi_{31}^c) + \left[ \left( \frac{e_{31} p_3}{\epsilon_{33}} - \lambda_p \right) \cdot \left( \int_{-(h_e/2+h_p)}^{-h_e/2} \frac{\partial \theta}{\partial x} b dz + \int_{h_e/2}^{h_e/2+h_p} \frac{\partial \theta}{\partial x} b dz \right) - \lambda \int_{-h_e/2}^{h_e/2} \frac{\partial \theta}{\partial x} b dz \right] \right\} \frac{\partial u_z}{\partial x} \\ & + \left\{ \tilde{K} \left[ \frac{\partial u_x}{\partial x} + \frac{1}{2} \left( \frac{\partial u_z}{\partial x} \right)^2 \right] + e_{31} b (\phi_{33}^c + \phi_{31}^c) + \left[ \left( \frac{e_{31} p_3}{\epsilon_{33}} - \lambda_p \right) \cdot \left( \int_{-(h_e/2+h_p)}^{-h_e/2} \theta b dz + \int_{h_e/2}^{h_e/2+h_p} \theta b dz \right) - \lambda \int_{-h_e/2}^{h_e/2} \theta b dz \right] \right\} \frac{\partial^2 u_z}{\partial x^2} = \bar{m} \ddot{u}_z . \end{aligned} \quad (55b)$$

Since the longitudinal inertia is negligible, i.e.,  $\bar{m} \ddot{u}_x \approx 0$ , factoring the partial derivatives and regrouping the force and moments gives

$$\frac{\partial N_x}{\partial x} = 0 , \quad (56)$$

$$\frac{\partial^2 M_x}{\partial x^2} + N_x \frac{\partial^2 u_z}{\partial x^2} = \bar{m} \ddot{u}_z . \quad (57)$$

Eq.(56) implies that the axial force  $N_x$  is not a function of  $x$ , i.e.,  $N_x = \text{constant}$ . Considering individual elastic, control, and temperature effects, one can further write Eq.(57) as

$$-\tilde{D} \frac{\partial^4 u_z}{\partial x^4} + \frac{\partial^2 M_x^c}{\partial x^2} + \frac{\partial^2 M_x^t}{\partial x^2} + N_x \frac{\partial^2 u_z}{\partial x^2} = \bar{m} \ddot{u}_z , \quad (58)$$

where  $N_x = \tilde{K} \left[ \frac{\partial u_x}{\partial x} + \frac{1}{2} \left( \frac{\partial u_z}{\partial x} \right)^2 \right] + N_x^c + N_x^t$ . The nonlinear piezothermoelastic beam equation can be further simplified when boundary conditions are specified. In the following two cases, one is used to compare with the standard equation and the other is for a detailed parametric study.

### 1) Free Expansion/Contraction:

If the longitudinal motion either at  $x = 0$  or at  $x = L$  is not constrained (free expansion/contraction), the axial force  $N_x$  vanishes when the boundary conditions are imposed. The differential equation then can be simplified to

$$-\tilde{D} \frac{\partial^4 u_z}{\partial x^4} + f(x,t) = \bar{m} \ddot{u}_z, \quad (59)$$

where  $f(x,t) = \frac{\partial^2 M_x^c}{\partial x^2} + \frac{\partial^2 M_x^t}{\partial x^2}$ . This is a standard form of the beam transverse vibration (Meirovitch, 1975). However, note that the physical meaning is much more complicated than the conventional form, due to the coupling of mechanical, electric, and temperature fields in the nonlinear piezothermoelastic laminated beam.

### 2) Simply Supported with Both Ends Fixed:

Boundary conditions for a simply supported piezothermoelastic laminated beam with both ends fixed are

$$u_z = u_x = 0, \text{ and } M_x = 0, \quad (60)$$

at both beam ends:  $x = 0$  and  $x = L$ . Furthermore, it is assumed the voltage  $\phi$  and temperature variation  $\theta$  are uniform in the  $x$ -direction. This implies that  $\phi$  and  $\theta$  are not functions of coordinate  $x$ . Then, the transverse equation becomes

$$-\tilde{D} \frac{\partial^4 u_z}{\partial x^4} + N_x \frac{\partial^2 u_z}{\partial x^2} = \bar{m} \ddot{u}_z \quad (61)$$

where the axial force  $N_x = \tilde{K} \left[ \frac{\partial u_x}{\partial x} + \frac{1}{2} \left( \frac{\partial u_z}{\partial x} \right)^2 \right] + N_x^c + N_x^t$ . Solution procedures of the



simply supported nonlinear piezothermoelastic beam equation are discussed next. Numerical results and control effectiveness are presented in case studies.

## CONTROL OF NONLINEAR DEFORMATION AND FREQUENCIES

As discussed previously, distributed piezoelectric layers laminated (coupled or embedded) on elastic shell continua can be used as distributed sensors and/or actuators (Tzou, 1991; Tzou, Zhong, and Natori, 1993; Tzou, Zhong, and Hollkamp, 1994). Injecting high voltages into the distributed piezoelectric actuators induces two major control actions. One is the in-plane membrane control force(s) and the other is the out-of-plane bending control moment(s) (Tzou, 1991;1993). In general, the control moments are essential in planar structures, e.g., plates and beams (Tzou and Fu, 1994); the membrane control forces are effective in shells (Tzou, Zhong, Hollkamp, 1994). In this study, the piezoelectric actuators are used to control the nonlinear large deformation and amplitude-dependent frequencies of flexible beams and plates, and their control effectivenesses are evaluated. General solutions of the nonlinear equations can be derived by a number of methods, e.g., the double Fourier series method, the Ritz method, Galerkin's method, the perturbation method, etc (Chia, 1980). In order to investigate the coupling among elastic, electric, temperature and control effects of the piezothermoelastic laminated beam, analytical solutions, including all design and control variables, are derived. The solution procedures are divided into two parts. The first step is to solve for nonlinear static solutions and the second step is to solve for dynamic solutions with respect to the nonlinear static equilibrium position. Numerical solutions are derived to evaluate the control effectiveness of nonlinear beams and plates in the case studies presented next.

## CASE STUDIES

### Case-1:

#### Control of Nonlinear Piezoelectric Laminated Plates

A simply supported nonlinear square plate (dimensions:  $2a \times 2b \times h$ ) is considered in the case study. The piezoelectric laminated plate is subjected to both thermal and electric bending loads:  $\phi_{33} = -\phi_{31} = \phi$  and  $\theta(z) = \Delta\theta(z/h)$  where  $\Delta\theta = \theta_t - \theta_b$  and  $\theta_t, \theta_b$  are the temperatures on the top and the bottom of the plate, respectively. Thus, the thermal and

the electric bending moment are

$$\begin{bmatrix} M_{xx}^{\theta} \\ M_{yy}^{\theta} \end{bmatrix} = \int_{-h/2}^{+h/2} \begin{bmatrix} \lambda_x \\ \lambda_y \end{bmatrix} \theta_z dz = \begin{bmatrix} \lambda_x \\ \lambda_y \end{bmatrix} \frac{\Delta \theta}{h} \int_{-h/2}^{+h/2} z^2 dz = \begin{bmatrix} \lambda_x \\ \lambda_y \end{bmatrix} \frac{h^2}{12} \Delta \theta, \quad (62)$$

$$\begin{bmatrix} M_{xx}^e \\ M_{yy}^e \end{bmatrix} = 2\phi e_{31} \Gamma_a \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \quad (63)$$

A nondimensional loading parameter  $\delta = \frac{2\phi e_{31} \Gamma_a (2a)^2}{Y h^4} + \frac{\theta_t \lambda (2a)^2}{6 Y h^2}$  is defined by the piezoelectric constant, the moment arm, Young's modulus, the thermal stress coefficient, the plate thickness and width, and, of course, the temperature and the control voltage. The central deflections due to the nondimensional loading  $\delta$  calculated based on the linear and nonlinear theories are plotted in in Figure 4. Convergence of the deflection solutions is sufficiently fast and the results obtained by a one-term approximation in the double-series expansion are adequate (Sundara, et. al., 1966).

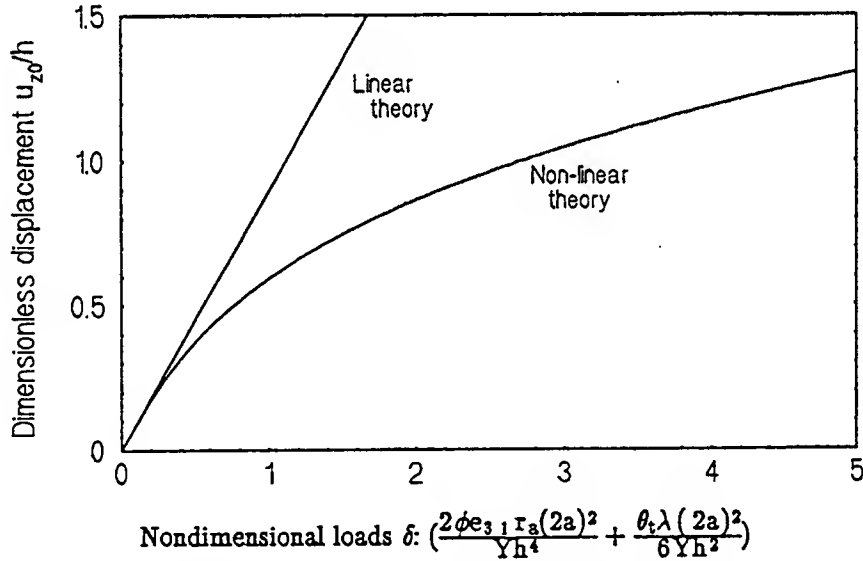


Fig.4 Central deflections of a simply supported piezoelectric laminated plate.

Center deflections of the simply supported nonlinear composite plate with control voltages and temperature loadings are analyzed and results are plotted in Figures 5–7. Note that the elastic plate is made of steels and the piezoelectric material layers are either PZT or PVDF materials. The plate dimensions are: steel thickness  $h = 1.0 \times 10^{-3} \text{m}$ , piezoelectric layer thickness  $h_p = 6.0 \times 10^{-5} \text{m}$ , plate length/width  $2a = 0.5 \text{m}$ .

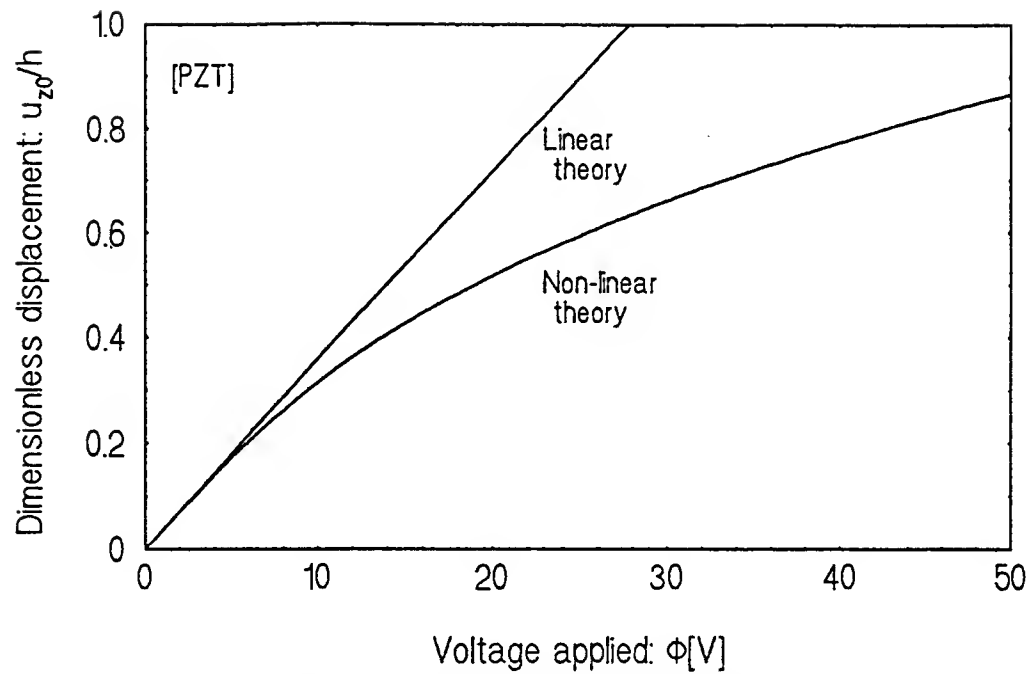


Fig.5 Central deflections of plate versus voltage applied (PZT).

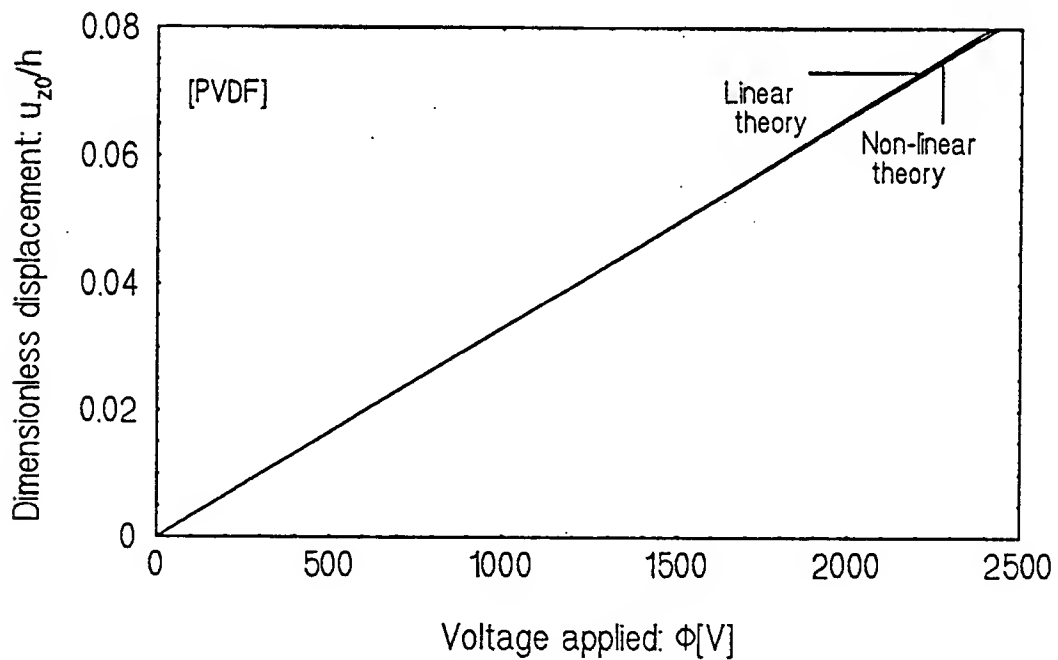
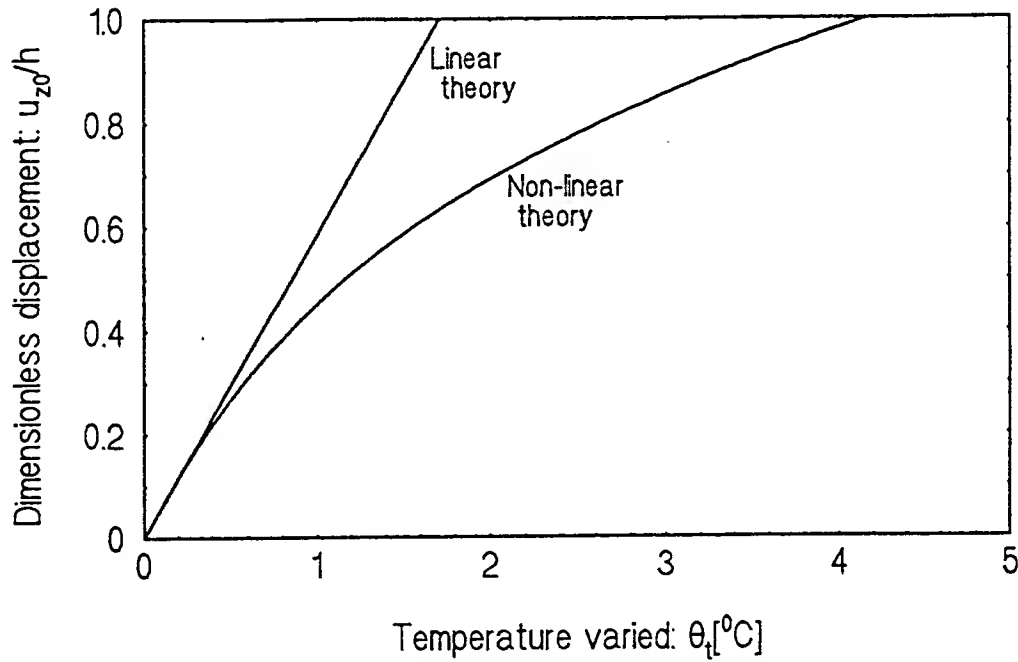


Fig.6 Central deflections of plate versus voltage applied (PVDF).



**Fig.7 Central deflections of plate versus temperature variation.**

Figures 5–6 show the linear and nonlinear relationships between the plate deflection and the control voltage. This voltage induced action can be used to counteract the deflections induced by the mechanical or temperature loadings, e.g., Figure 7. Besides, the PZT induced control action is superior to the PVDF induced control action. (This can be easily inferred from the inherent piezoelectric constants:  $e_{31} = 10.43 \text{ C/m}^2$  for PZT while  $e_{31} = 9.6 \times 10^{-3} \text{ C/m}^2$  for PVDF.)

#### **Case-2:**

##### **Control of Nonlinear Piezoelectric Laminated Beams**

It is assumed that a simply supported three layer PZT/steel/PZT beam with dimensions: width  $b = 0.0508\text{m}$ , length  $L = 1\text{m}$ , steel thickness  $h_e = 0.00635\text{m}$ , and lead zirconate titanate (PZT) thickness  $h_p = 254 \times 10^{-6}\text{m}$  is used in the case study, Figure 8. Detailed material properties are summarized in Appendix: Table 1 and Table 2.

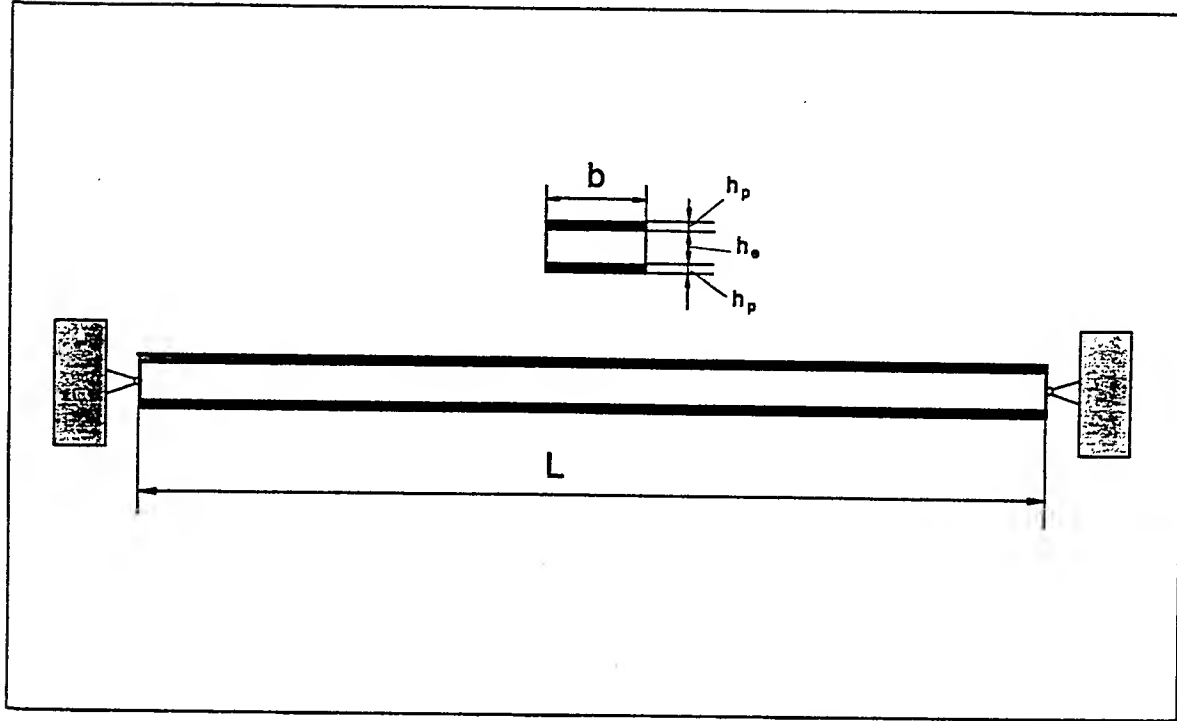


Fig.8 A PZT/steel/PZT laminated beam.

Next, the bending stiffness  $\tilde{D}$  and the membrane stiffness  $\tilde{K}$  can be respectively calculated as:  $\tilde{D} = YI + 2Y_p I_p + 2I_p e_{31}^2 / \epsilon_{33} = 93.765(\text{N} \cdot \text{m}^2)$ ;  $\tilde{K} = Y\tilde{A} + 2Y_p \tilde{A}_p + 2\tilde{A}_p e_{31}^2 / \epsilon_{33} = 23.986 \times 10^6(\text{N})$ . Note that the values of  $YI$ ,  $2Y_p I_p$ , and  $2I_p e_{31}^2 / \epsilon_{33}$  in the PZT/steel/PZT beam are 80%, 18%, and 2% of the total bending stiffness  $\tilde{D}$ , respectively, and values of  $Y\tilde{A}$ ,  $2Y_p \tilde{A}_p$ , and  $2\tilde{A}_p e_{31}^2 / \epsilon_{33}$  are 92.8%, 6.5%, and 0.7% of the total membrane stiffness  $\tilde{K}$ . It is assumed that applied control voltages  $\phi_{3\ddot{\zeta}}$  and  $\phi_{3\dot{\zeta}}$  are uniformly distributed and  $\phi_{3\dot{\zeta}} = -\phi_{3\ddot{\zeta}} = \phi$ , and the temperature rise  $\theta$  is also uniform along the x-axis and linear variation through the thickness:  $\theta(z) = \bar{a}z + \bar{c}$ , where  $\bar{a} = (\theta_t - \theta_b) / (h_e + 2h_p)$ ,  $\bar{c} = (\theta_t + \theta_b) / 2$ ,  $\theta_t$  is the top surface temperature and  $\theta_b$  is the bottom surface temperature of the beam. Note that  $\theta_b = -\theta_t = \theta$  which implies that the total temperature difference between the top and bottom surfaces is  $2\theta$ . Then, the electric control bending moment  $M_x^c$  and the temperature induced moment  $M_x^t$  are

$$M_x^c = 0.003499\phi (\text{N} \cdot \text{m}) \quad \text{and} \quad M_x^t = 0.34856\theta (\text{N} \cdot \text{m}), \quad (64a,b)$$

in which 98% is due to the steel and only 2% is due to the PZT in the temperature induced bending moment. (Recall that the ceramics are less sensitive to temperatures as compared with steels.) The force and moment relationship can be simplified to

$$\left(\frac{\tanh v}{v} - \frac{1}{\cosh^2 v}\right) = \frac{64 v^4 \tilde{D}^3}{\tilde{K}(M_x^c + M_x^t)^2 L^4}. \quad (65)$$

Denoting  $y_1 = (\tanh v/v - 1/\cosh^2 v)$  and  $y_2 = 64v^4 \tilde{D}^3 / [\tilde{K}(M_x^c + M_x^t)^2 L^4]$ , one can plot  $y_1(v)$  and  $y_2(v)$ . Intersections of  $y_1(v)$  and  $y_2(v)$  gives solutions  $v$  of Eq.(65), such as shown in Figures 9 to 11. Then, the axial force  $N_{x,s}$  and the beam center deflection (at  $x = L/2$ ) can be calculated and its temperature/control effects studied.

$$N_{x,s} = \frac{4v^2 \tilde{D}}{L^2}, \quad (66)$$

$$u_{z,s} \Big|_{x=L/2} = \frac{(M_x^c + M_x^t)L^2}{4\tilde{D}v^2} \left( \frac{1}{\cosh v} - 1 \right). \quad (67)$$

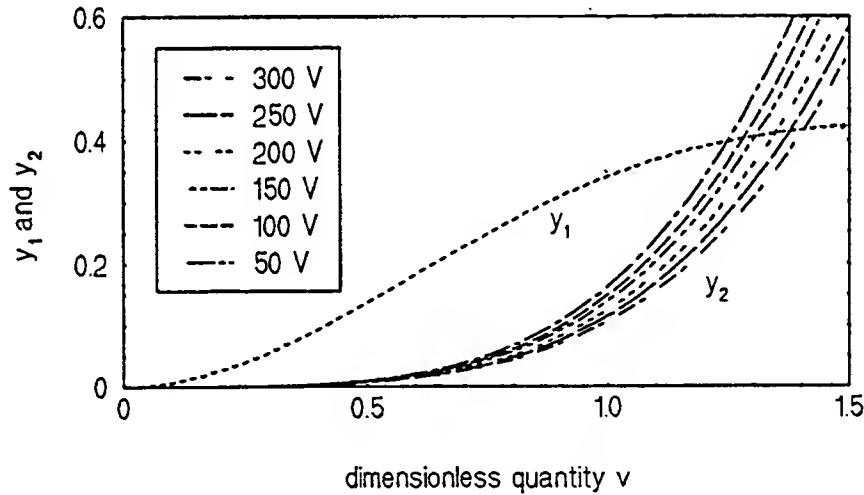


Fig.9 Solution  $v$  for various control voltages (temperature  $\theta = 10^\circ \text{C}$ ).

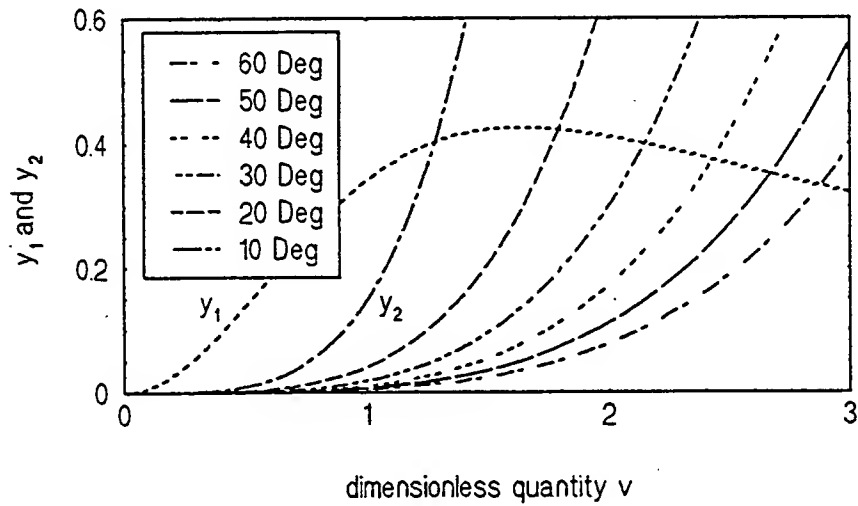


Fig.10 Solution  $v$  for various temperatures (voltage  $\phi = 100V$ ).

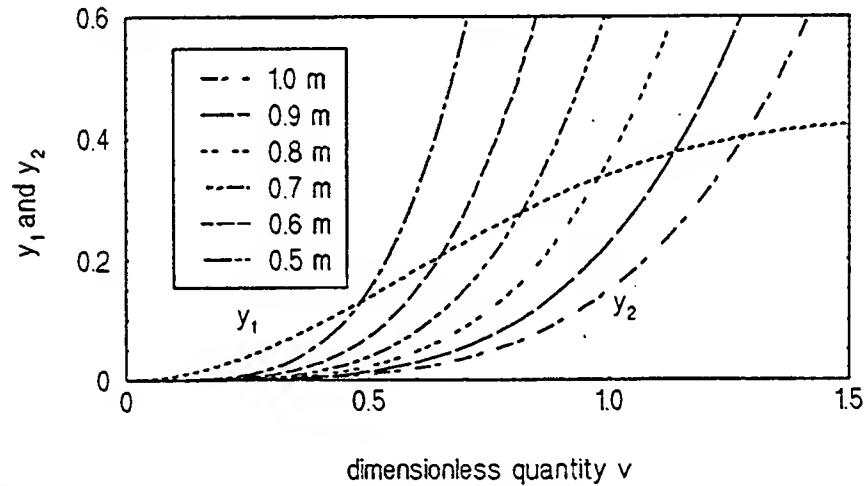
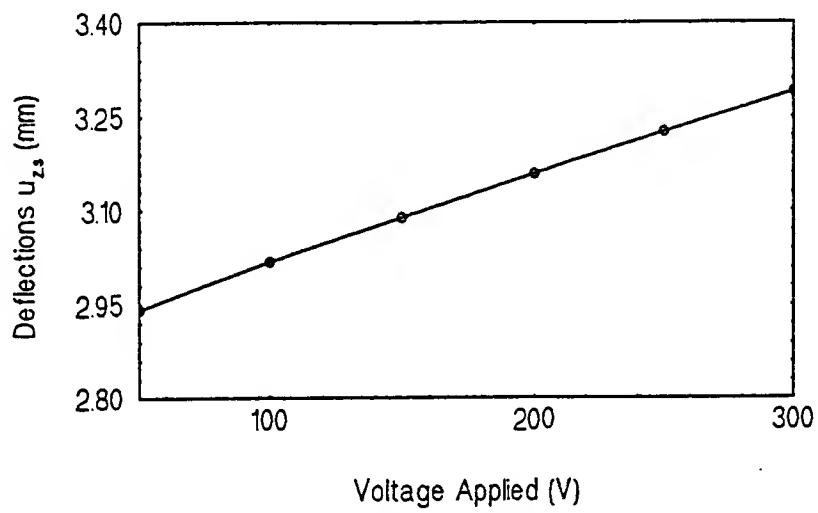
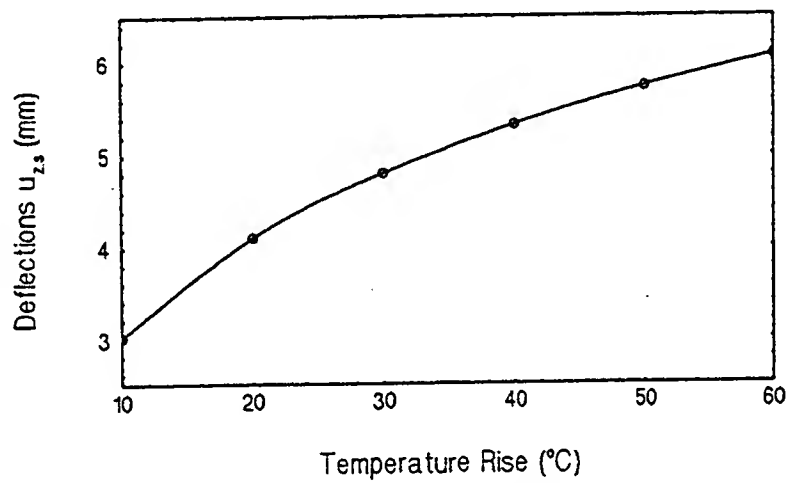


Fig.11 Solution  $v$  for various beam lengths ( $\theta = 10^\circ C$  and  $\phi = 100V$ ).

Static deflections of the beam center ( $x = L/2$ ) with respect to the applied control voltage (at  $\theta = 10^\circ C$ ), temperature rise (at  $\phi = 100V$ ), and beam length (with  $\phi = 100V$ ,  $\theta = 10^\circ C$ ) are plotted in Figures 12–14. Note that the  $10^\circ C$  temperature represents a total of  $20^\circ C$  difference between the top and bottom surfaces. The deflection and voltage relation, Figure 12, gives a general guideline that the control voltage induced displacement can be used to compensate the temperature induced deflection or the nonlinear deflection. Equivalent axial force with respect to the beam center deflection is presented next.



**Fig.12** Static deflections v.s. voltage changes (temperature  $\theta = 10^\circ \text{C}$ ).



**Fig.13** Static deflections v.s. temperature changes (voltage  $\varphi = 100\text{V}$ ).



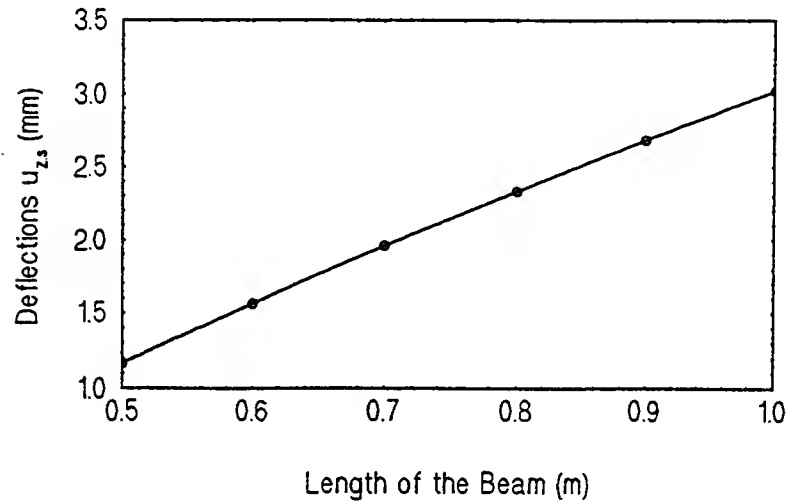


Fig.14 Static deflections for various beam lengths ( $\theta = 10^\circ \text{C}$  and  $\varphi = 100\text{V}$ ).

Figure 15 shows the axial force versus the static deflections of the beam center, which reveals that the induced axial control force stiffens the beam and consequently the natural frequencies of the beam increase. (Note that this force can also be viewed as an axial control force.) The frequency increase can be expressed by the quantity  $\{[1 + (N_{x,s} L^2 / i^2 \bar{D} \pi^2)]^{1/2} - 1\} \times 100$  percent, where  $i$  is the mode number, and the results are shown in Figures 16–18. The percentage of variation for the first mode is higher than those of the higher modes. The numerical results suggest that both static deflection and dynamic behaviors of the simply supported nonlinear PZT/steel/PZT laminated beam are influenced by the temperature and they also can be controlled by the control voltages applied to the piezoelectric actuators.

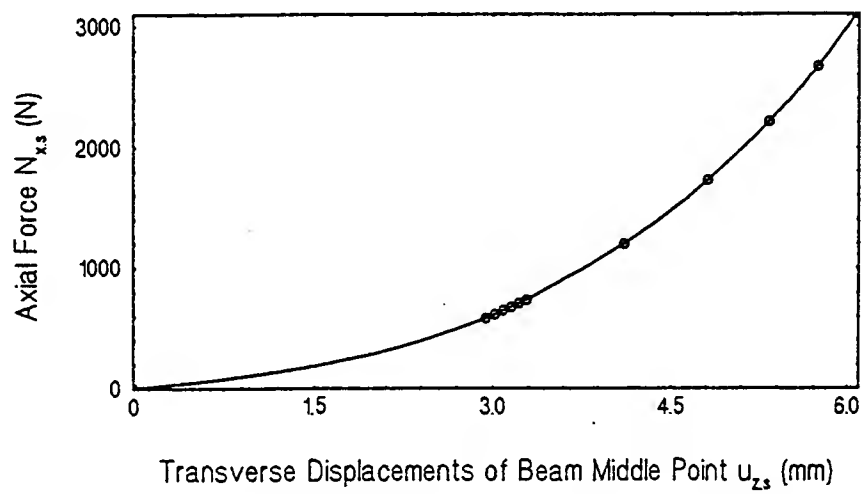


Fig.15 Axial forces v.s. beam deflections.

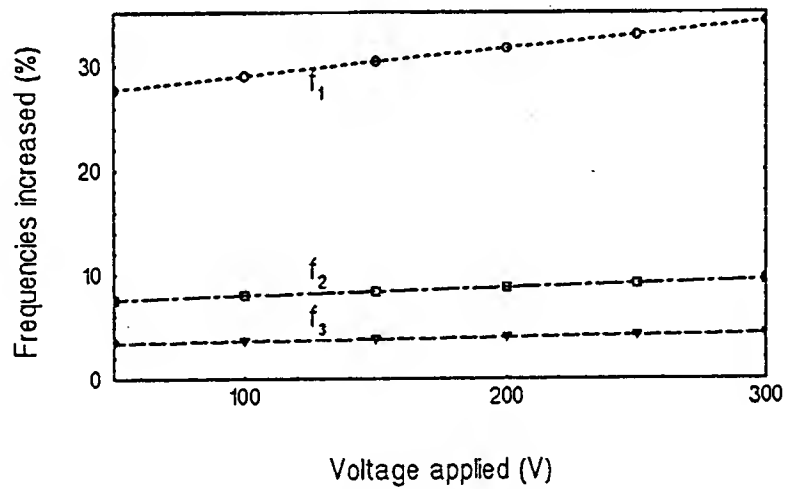


Fig.16 Frequency variations v.s. control voltages ( $\theta = 10^\circ \text{C}$ ).

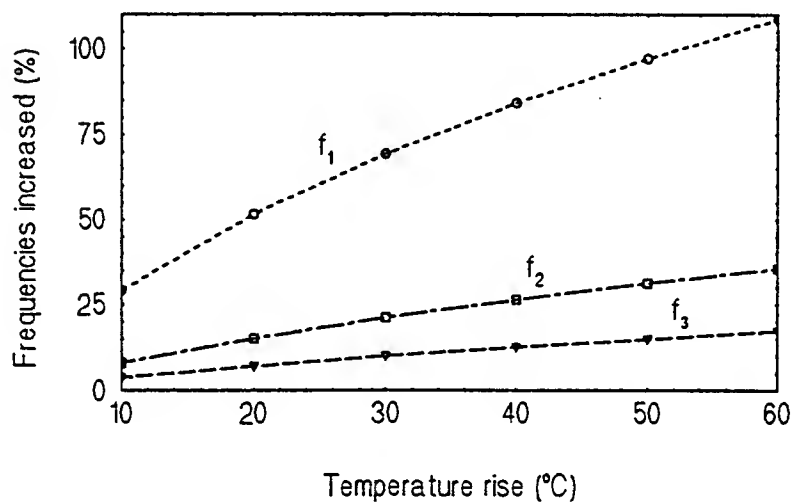


Fig.17 Frequency variations v.s. temperatures ( $\phi = 100$  V).

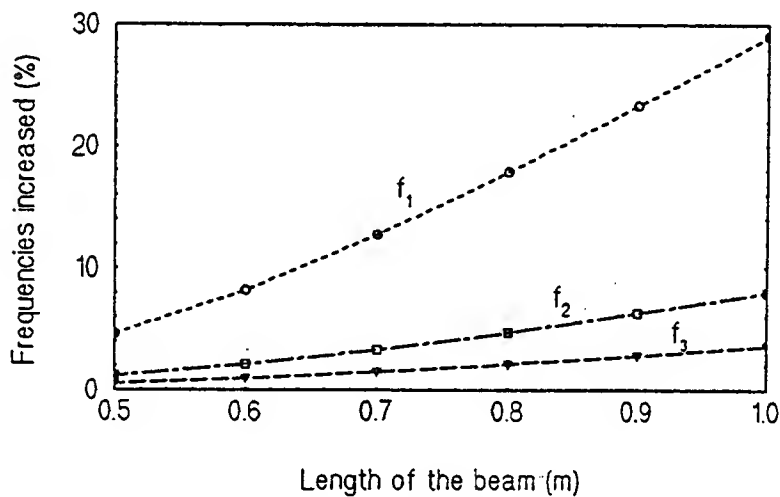


Fig.18 Frequency variations v.s. beam lengths ( $\theta = 10^\circ$  C,  $\phi = 100$  V).

## SUMMARY AND CONCLUSIONS

Beam and plate like structures and components are widely used in many aerospace structures. Imposed shape changes and surface control of flexible beams and plates could offer many aerodynamic advantages in flight maneuverability and precision control. Deformed shapes and surfaces often involve nonlinear deformations. Studies of control of nonlinear behavior related to the deformed surfaces and shape changes would provide detailed information in future controlled surface design and implementation. This research

is concerned with the control effectiveness of nonlinearly deformed beams and plates based on the smart structures technology.

In the recent development of smart structures and structronic systems, piezoelectric materials are widely used as sensors and actuators in sensing, actuation, and control applications. This research is to investigate the control effectiveness of piezoelectric laminated nonlinear flexible beams and plates (rectangular and square plates) subjected to mechanical and temperature excitations. It is assumed that the flexible beams and plates encounter the von Karman type geometrical nonlinearity. Thermoelectromechanical equations and boundary conditions including elastic, temperature, and piezoelectric couplings will be formulated first, and analytical solutions be derived. The reduced nonlinear static equations are identical to the classic nonlinear plate and beam equations. Dynamics, electromechanical couplings, and control of nonlinear piezoelectric laminated beams and plates with large deformations are investigated. Active control of nonlinear flexible deflections, thermal deformations, and natural frequencies using distributed piezoelectric actuators are studied, and their nonlinear effects are evaluated.

Nonlinear static deflections with the influence of temperature and control voltage were studied. Voltage-temperature and displacement relations for piezoelectric laminated nonlinear plates and beams were investigated. The voltage imposed actuations can be used to compensate the nonlinear deformation and the temperature induced deformation. Small-amplitude beam oscillations with respect to the nonlinearly deformed static equilibrium position were investigated. It was observed that the total bending stiffness of the PZT/steel/PZT laminated beam is contributed by the steel (80%) and PZT (elasticity: 18% and piezoelectricity: 2%), and the total membrane stiffness is contributed by the steel (98%) and PZT (elasticity: 6.5% and piezoelectricity: 0.7%) in the laminated beam. The piezoelectricity contributed stiffness is relatively insignificant. Simulation results also suggested that the voltage induced control displacement/force can be used to compensate the nonlinear static deflection, temperature effects, and natural frequencies of the piezoelectric laminated beam.

## ACKNOWLEDGEMENT

This research is supported, in part, by a grant (96-0882)(F49620-93-0063) from the Air Force Office of Scientific Research and the Wright Laboratory (Flight Dynamics Laboratory). Technical advices from Dr. V.B. Venkayya are gratefully acknowledged. This report constitutes a part of Dr. Y. Bao's Ph.D. dissertation.

## BIBLIOGRAPHY

Chia, C.Y., 1980, *Nonlinear Analysis of Plates*, McGraw-Hill International Book Co., New York.

Lalande, F., Chaudhry, Z., and Rogers, C.A., 1993, "A Simplified Geometrically Nonlinear Approach to the Analysis of the Moonie Actuator," *Adaptive Structures and Material Systems*, AD-Vol.35, pp.149-155, 1993 ASME WAM, New Orleans, LA, Nov.28-Dec.3, 1993.

Librescu, L., 1987, "Refined Geometrically Nonlinear Theories of Anisotropic Laminated Shells," *Quarterly of Applied Mathematics*, Vol.XLV, No.1, pp.1-22.

Meirovitch, L., 1975, *Elements of Vibration Analysis*, McGraw-Hill, p.208.

Palazotto, A.N. and Dennis, S.T., 1992, *Nonlinear Analysis of Shell Structures*, AIAA Pub., Washington, D.C.

Pai, P.F., Nafeh, A.H., Oh, K., and Mook, D.T., 1993, "A Refined Nonlinear Model of Piezoelectric Plate," *J. Solids and Structures*, Vol.30, pp.1603-1630.

Pietraszkiewicz, W., 1979, *Finite Rotations and Lagrangean Description in the Non-linear Theory of Shells*, Polish Academy of Science, Polish Scientific Publishers.

Soedel, W., 1981, *Vibrations of Shells and Plates*, Dekker, New York, p.271.

Sreeram, P.N., Salvady, G. and Naganathan, N.G., 1993, "Hyteresis Prediction for a Piezoceramic Material System," *Adaptive Structures and Material Systems*, AD-Vol.35, pp.35-42, 1993 ASME WAM, New Orleans, LA, Nov.28-Dec.3, 1993.

Timoshenko, S., Young, D.H., and Weaver, W., 1974, *Vibration Problems in Engineering*, Wiley, New York.

Tzou H.S. and Anderson, G.L., (Ed.), 1992, *Intelligent Structural Systems*, Kluwer Academic Publishers, Dordrecht/Boston/London.

Tzou, H.S., 1993, *Piezoelectric Shells (Distributed Sensing and Control of Continua)*, Kluwer Academic Publishers, Dordrecht/Boston/London.

Tzou, H.S. and Bao, Y., 1994, "Modeling of Thick Anisotropic Composite Triclinic Piezoelectric Shell Transducer Laminates," *Journal of Smart Materials and Structures*, Vol.3, pp.285–292.

Tzou, H.S. and Bao, Y., 1995, "A Theory on Anisotropic Piezothermoelastic Shell Laminates with Sensor/Actuator Applications," *Journal of Sound & Vibration*, Vol.184, No.3, pp.453–473.

Tzou, H.S. and Bao, Y., 1996, "Nonlinear Piezothermoelasticity and Multi-field Actuations, Part-1: Nonlinear Anisotropic Piezothermoelastic Shell Laminates," *ASME Transactions, Journal of Vibration & Acoustics*. (To appear)

H.S. Tzou and T. Fukuda (Ed), 1992, *Precision Sensors, Actuators, and Systems*, Kluwer Academic Publishers, Dordrecht/Boston/London, August 1992.

Tzou, H.S. and Howard, R.V., 1994, "A Piezothermoelastic Thin Shell Theory Applied to Active Structures," *ASME Transactions, Journal of Vibration & Acoustics*, Vol.116, No.3, pp.295–302.

Tzou, H.S. and Ye, R., 1994, "Piezothermoelasticity and Precision Control of Piezoelectric Systems: Theory and Finite Element Analysis," *ASME Transactions, Journal of Vibration & Acoustics*, Vol.116, No.4, pp.489–495.

Tzou, H.S. and Zhong, J.P., 1993, "Electromechanics and Vibrations of Piezoelectric Shell Distributed Systems: Theory and Applications," *ASME Journal of Dynamic Systems, Measurements, and Control*, Vol.115, No.3, pp.506–517.

Tzou, H.S., Zhong, J.P., and Natori, M.C., 1993, "Sensor Mechanics of Distributed Shell Convolver Sensors Applied to Flexible Rings," *ASME Journal of Vibration & Acoustics*, Vol.(115), No.1, pp.40–46, January 1993.

Tzou, H.S., Zhong, J.P., and Hollkamp, J.J., 1994, "Spatially Distributed Orthogonal Piezoelectric Shell Actuators: Theory and Applications," *Journal of Sound & Vibration*, Vol.177, No.3, pp.363–378, October 1994.

Tzou, H.S. and Zhou, Y-H., 1995, "Dynamics and Control of Nonlinear Circular Plates with Piezoelectric Actuators," *J of Sound & Vibration*, Vol.188, No.2, pp.189–207.

Yu, Y.Y., 1993, "Some Recent Advances in Linear and Nonlinear Dynamical Modeling of Elastic and Piezoelectric Plates," *Adaptive Structures and Material Systems*, AD-Vol.35, pp.185–195, 1993 ASME WAM, New Orleans, LA, Nov.28–Dec.3, 1993.

## APPENDIX: MATERIAL PROPERTIES

Table 1 PZT material properties.

Young's modulus	$Y_x = Y_y = Y_z = 61 \text{ GPa}$
Shear modulus	$G_{xy} = G_{xz} = G_{yz} = 23.64 \text{ GPa}$
Poisson's ratio	$\mu = 0.29$
Density	$\rho = 7.7 \times 10^3 \text{ kg/m}^3$
Thermal expansion coefficient	$\alpha = 1.2 \times 10^{-6} \text{ m/m/}^\circ\text{C}$
Thermal stress coefficient	$\lambda_p = 1.03 \times 10^5 \text{ N/m}^2/^\circ\text{C}$
Electric permittivity	$\epsilon_{33} = 1.65 \times 10^{-8} \text{ F/m}$
Piezoelectric constant	$d_{31} = 171 \times 10^{-12} \text{ C/N (m/V)}$ $e_{31} = 10.43 \text{ C/m}^2$
Pyroelectric constant	$p_3 = 0.25 \times 10^{-4} \text{ C/m}^2/^\circ\text{C}$

Table 2 Steel material properties.

Young's modulus	$Y_x = Y_y = Y_z = 68.95 \text{ GPa}$
Shear modulus	$G_{xy} = G_{xz} = G_{yz} = 26.52 \text{ GPa}$
Poisson's ratio	$\mu = 0.30$
Density	$\rho = 7.75 \times 10^3 \text{ kg/m}^3$
Thermal expansion coefficient	$\alpha = 1.1 \times 10^{-5} \text{ m/m/}^\circ\text{C}$
Thermal stress coefficient	$\lambda = 1.08 \times 10^6 \text{ N/m}^2/^\circ\text{C}$

(Rdl-Rept97.Wp/Fi197.t3b1)

# A Progressive Refinement Approach to Planning and Scheduling

William J. Wolfe  
Associate Professor  
Department of Computer Science and Engineering

University of Colorado at Denver  
Denver, CO

Final Report for:  
Summer Research Extension Program  
Wright Laboratory

Sponsored by:  
Air Force Office of Scientific Research  
Bolling AFB, Washington, DC

and

Wright Laboratory

January 1997



# A Progressive Refinement Approach to Planning and Scheduling

William J. Wolfe

Associate Professor of Computer Science  
Department of Computer Science and Engineering  
University of Colorado at Denver

## Abstract

A summary of the algorithmic results presented in this report: 1. We introduce a "novel" scheduling problem called the window-constrained packing problem (WCP), and we provide extensive simulation results; 2. We describe the " $\theta$ -filter" approach, a creative way to deal with multiple sorting features. 3. The "priority dispatcher" that we invented uses three phases: i. selection/sort; ii. allocation; iii. optimization. This simple modular design can be modified in a variety of ways, making it easy to adapt to specific applications; 4. The "look-ahead" algorithm that we invented uses the dispatcher to look ahead in the list of jobs to determine a good choice of allocation options; 5. The "augmented genetic" algorithm that we invented has the potential to produce near-optimal results, when time permits.

There are so many applications that require advanced planning and scheduling techniques that we can only provide a partial listing, such as: Transportation Systems (Airlines, Trucking, Logistics, Buses, Highways, etc.); Facilities Management (Hospitals, Courts, Schools, etc.); Communications Systems (Telephones, Cable, etc.); Factories; Spacecraft (Earth Observing, Deep Space, Communications, etc.). We identified three factors for evaluating the many algorithmic approaches: speed, simplicity, and accuracy. "Speed" refers to the time it takes an algorithm to process and return its results. "Simplicity" refers to the ease by which users are able to understand the algorithm's logic. "Accuracy" refers to the quality of the algorithm's results as measured by some criteria (i.e.: an objective function). The complexity of these environments pushed us toward fast and simple algorithms, such as *dispatch* methods. For increased accuracy we developed a *look ahead* method (relatively fast, relatively simple, and very accurate). If the accuracy requirements are severe (i.e.: need near-optimal solutions) then simplicity and speed can be sacrificed in favor of more complex and time consuming methods (e.g.: simulated annealing, genetic, tabu search, etc.). We have concluded that look-ahead approaches, and variations on that theme, are very often the best compromise, usually providing high quality results within reasonable run time limits, while also being relatively easy to understand. It is very difficult, and in many cases impossible, to define a reliable objective function. Therefore in most cases there is no easy way of comparing algorithmic results other than to say that one performed better with respect to *that* criteria. This observation pointed us toward fast and simple methods, since more complex methods are, more or less, using a lot of processing time to fine-tune the results according to a *presumed* objective function. Thus, there are at least three reasons why the pursuit of accuracy is questionable: 1. accuracy is often defined by a subjective weighting of multiple, disparate, criteria (weighing *apples* and *oranges*); 2. the time between planning and implementation is often long enough for many of the assumptions the plan is based on to become invalid; and 3. computing accurate solutions often takes a prohibitive amount of processing time. There is little point in pursuing accuracy at the expense of speed and simplicity in a rapidly changing environment. Highly accurate algorithmic results can be made obsolete by small changes in the constraints. The constraints themselves are often vague or imprecise. Additionally, we do not want a high speed scheduler that reacts to every little change, constantly moving jobs around. The scheduling process could become "unstable": rapidly changing commitments (confused customers and operators). These considerations led us to incorporate a "fuzzy" approach. In fact, we have tentatively concluded that highly complex, dynamic, domains with multiple, possibly vague, criteria are prime application areas for the fuzzy logic approach. To add further context to these algorithmic results we also explored a hierarchical approach called "progressive refinement". It takes into account a *rolling horizon* of *out-day* schedules that are continuously adapted until operational time. This approach helps to smooth over many of the de-stabilizing inputs (e.g.: machine failure, canceled orders, new orders, price changes, etc.).

# 1. Introduction

---

The development of automated scheduling systems continues to be an area of great interest in industry and academia. It involves the computer modeling of *jobs*, *resources*, and *constraints*, as well as the design of *algorithms* and other systems that would help human schedulers do their jobs. There are so many applications that require advanced planning and scheduling techniques that we can only provide a partial listing, such as:

- Transportation Systems (Airlines, Trucking, Logistics, Buses, Highways, etc.);
- Facilities Management (Hospitals, Courts, Schools, etc.);
- Communications Systems (Telephones, Cable, etc.);
- Factories;
- Spacecraft (Earth Observing, Hubble Space Telescope, Deep Space, Communications, etc.).

We identified three factors for evaluating the many algorithmic approaches: speed, simplicity, and accuracy. "Speed" refers to the time it takes an algorithm to process and return its results. "Simplicity" refers to the ease by which users are able to understand the algorithm's logic. "Accuracy" refers to the quality of the algorithm's results as measured by some criteria (i.e.: an objective function).

The application areas mentioned above are very complex and highly dynamic. The complexity of these environments pushed us toward fast and simple algorithms, such as *dispatch* methods. For increased accuracy we developed a particular *look ahead* method (relatively fast, relatively simple, and very accurate). If the accuracy requirements are severe (i.e.: need near-optimal solutions) then simplicity and speed can be sacrificed in favor of more complex and time consuming methods (e.g.: simulated annealing, genetic, tabu search, etc.). We have concluded that look-ahead approaches, and variations on that theme, are very often the best compromise, usually providing high quality results within reasonable run time limits, while also being relatively easy to understand.

Because of the complexity of these domains it is very difficult, and in many cases impossible, to define a reliable objective function. Therefore in most cases there is no easy way of comparing algorithmic results other than to say that one performed better with respect to *that* criteria. This observation also points us toward fast and simple methods, since more complex methods are, more or less, using a lot of processing time to fine-tune the results according to a *presumed* objective function.

Thus, there are at least three reasons why the pursuit of "accuracy" is questionable: 1. accuracy is often defined by a subjective weighting of multiple, disparate, criteria (weighing *apples* and *oranges*); 2. the time between planning and implementation is often long enough for many of the assumptions the plan is based on to become invalid; and 3. computing accurate solutions often takes a prohibitive amount of processing time. There is little point in pursuing accuracy at the expense of speed and simplicity in a rapidly changing environment (e.g.: rapidly changing goals, priorities, cost factors, machine availability, etc.). Highly accurate algorithmic results can be made obsolete by small changes in the constraints. The constraints themselves are often vague or imprecise, further discouraging the fine-tuning of algorithmic results.

Additionally, we do not want a high speed scheduler that reacts to every little change, constantly moving jobs around. The scheduling process could become "unstable": rapidly changing commitments (confused customers, suppliers, and workers). These considerations led us to incorporate a "fuzzy" approach (section

6). In fact, we have tentatively concluded that highly complex, dynamic, domains with multiple, possibly vague, criteria are prime application areas for the fuzzy logic approach.

To add further context to these algorithmic results we also explored a hierarchical approach called "progressive refinement". It takes into account a *rolling horizon* of *out-day* schedules that are continuously adapted until operational time. This approach helps to smooth over many of the de-stabilizing inputs (e.g.: machine failure, canceled orders, new orders, price changes, etc.). Although many of the simulations presented in this report are based on a "batch" approach (i.e.: dispatch algorithms) it is clear that such methods must be integrated into progressive refinement, or *incremental*, methods in most applications.

The algorithms that we discuss in this report can be classified as follows.

Algorithm Classifications	
<p><b>Optimal:</b></p> <p>Depth First (exhaustive search)</p> <p>Branch and Bound</p> <p>Linear Programming</p>	<p><b>Heuristic:</b></p> <p>Constructive Heuristics: Priority Dispatcher Look Ahead Fuzzy Logic</p> <p>Improvement Heuristics: Hill Climbing Simulated Annealing Tabu Search</p> <p>Repair Heuristics: Iterative Refinement</p> <p>Genetic: Indirect Representation Augmented Genetic.</p>

*Optimal* algorithms are guaranteed to find optimal solutions, but they usually take a prohibitive amount of time. The *constructive* heuristics begin with a batch of jobs and a *clean slate*, and schedule the jobs one by one (with limited backtracking) until a complete schedule is achieved. The *improvement* heuristics begin with a schedule and find ways to improve the quality. A constructive heuristic is often used to create a schedule that is then used as the starting point for an improvement heuristic.

The *repair* heuristics apply when a sudden "disturbance" (e.g.: machine failure, canceled order, etc.) has been introduced and a quick fix is needed. For example, the schedule might be amended in such a way as to cause the least amount of change to existing commitments. There are many variations on this theme. For example, a "schedule" might have many constraint violations (e.g.: overbooking) while it is going through a process of incremental improvement. Some authors call this "iterative refinement" [3]. One distinction that could be made is whether or not the momentary "schedules" are feasible<sup>1</sup> or can contain constraint violations. Until we get to the section on *progressive refinement* we assume that a "schedule" has no constraint violations.

The *genetic* approach is in a class by itself. It works by having a population of schedules and breeds new schedules from "pieces" of the best schedules in the current population.

---

<sup>1</sup> "Feasible" means that the schedule can be implemented (but it may be very inefficient).

We have found that we can begin with a particular heuristic approach and then create a variety of *hybrids* that use parts and pieces from any or all of the above methods. Thus, it is sometimes pointless to compare the individual methods since most applications will require a unique hybrid approach. We worked at understanding each method well enough to be able to quickly evaluate the cost-benefit trade off of including some part of the method in a hybrid approach to a real problem.

There are several issues that come up in one form or another throughout this report: algorithm processing time; algorithm rationale (i.e.: simplicity); multiple, conflicting, criteria; constraint propagation and relaxation; bottlenecks and other features of the job conflicts; general rules of thumb; uncertainty vs. complexity; dynamic environments; scheduling horizons; stability; progressive refinement.

**Overview of Planning and Scheduling concepts:** From allocating machines to jobs in a factory, to making lane changes on a highway, a wide variety of engineering problems involve some form of planning or scheduling: the allocation of resources to jobs over time. The term "planning" usually relates to higher level issues, such as the overall purpose, goals and strategies of an endeavor, while the term "scheduling" usually relates to lower level issues, such as specific equipment and operation start/stop times. In the simplest possible terms, *planning* deals with "what" and *scheduling* with "when". This distinction is usually minor and most authors, like us, use the terms interchangeably, but scheduling strongly connotes the specific assignment of *activities to time lines*, which is the main topic of this report.

There are several issues that make automated scheduling difficult, the primary ones being:

- *Ambiguity*: vague or poorly defined terms, criteria, constraints or specifications.
- *Modeling*: it is difficult to build efficient, high fidelity, models of the domain.
- *Uncertainty*: unexpected events and inherently random processes.
- *Combinatorial Explosion*: a huge number of alternatives and options.
- *Dynamics*: rapidly changing goals, resources, costs, requirements, etc.

Ambiguous criteria make it impossible to objectively tell when one schedule is better than another. For example, criteria such as "customer satisfaction" or "high quality product", are commonly expressed but are also difficult to quantify. They summarize a complex synthesis of several factors (possibly including such amorphous factors as the customer's *perceptions* of quality, etc.). The criteria can sometimes be clarified by identifying the contributing *components*, such as inventory cost, # of late jobs, machine preferences, average job flow time, etc. But individual components can conflict, for example: *low cost* and *on time* usually conflict with *high quality*<sup>2</sup>. And it may be very difficult to agree on *weighting factors* that might quantify the relative merits of trading off one criteria in favor of another (e.g.: is a little *lateness* worth a small increase in *quality*). The result of a successful analysis of the relevant criteria is a clearly specified *objective function*. The function could then be used to reliably distinguish not only that one schedule is *better* than another but also by *how much* it is better. Unfortunately, more often than not such an objective function is either a gross over simplification of the problem or based on several subjective parameters.

Criteria specifications can be related to at least these three broad categories: 1. Cost and profit factors; 2. Managerial policy decisions; 3. Customer preferences. Policy decisions and customer preferences are related to *profit*, but in a *long term* sense. Clearly, one major difficulty in quantifying the criteria is that it may change suddenly (market fluctuations, manager decisions, customer preferences, etc.).

---

<sup>2</sup> This reminds me of the sign that a student said was posted on a production manager's desk. It listed the following three items: "1. ON TIME; 2. HIGH QUALITY; 3. CHEAP"; under which was the advice: "PICK ANY TWO".

*Constraints* must also be clearly specified. There are at least three types of constraints: 1. Criteria-related; 2. Physical; 3. Both. Criteria specifications are sometimes interpreted as constraints. Take for example: "get all the jobs done on time". This is not a physical constraint, since it can be violated, but at a cost. Physical constraints are usually the kind that cannot be violated, such as size, shape, weight, or rate limitations related to the use of a particular piece of equipment. Capacity limits, exclusive use, precedence, and maximum number of shared users are also examples of physical constraints.

Some constraints are just *preferences* related to physical attributes of a product, such as color or size. Another way of classifying constraints is to distinguish *hard* from *soft* constraints. A hard constraint cannot be violated but a soft constraint has *degrees* of acceptability. For example, a soft constraint might be expressed as: the customer wants *green*, but *blue* is acceptable, and other colors are less acceptable. But the fact that the product must be painted might be a hard constraint. Now, it is also clear that when the boss says that there are to be no late jobs, then for all intents and purposes this is a hard constraint. The important distinction, however, is the difference between *relaxable* and *non-relaxable* constraints. For example, when the boss concludes that it is not possible for all the jobs to be on time, the due date constraint might be *relaxed* to allow jobs to be up to 1 day late. But the boss cannot dictate the violation of the laws of physics (e.g.: equipment operational limits). The variations along these lines are almost endless.

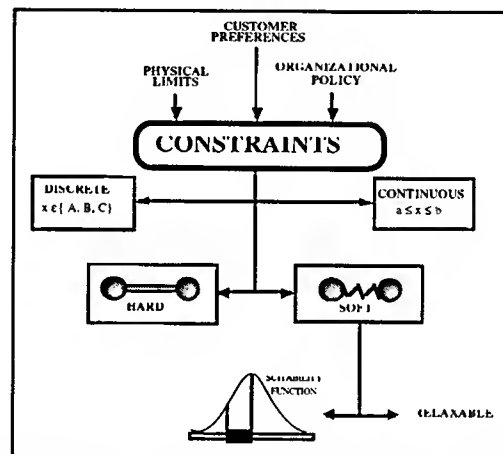
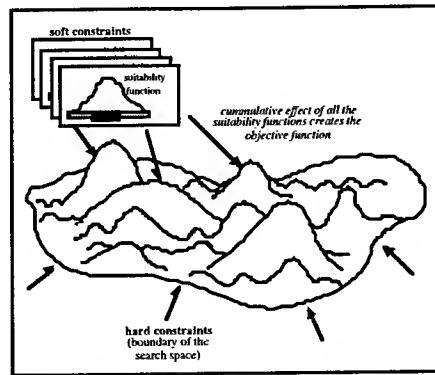


Figure A: There are several ways to classify constraints.

A last example: if there are only 5 machines in a shop then a valid schedule would use 5 or less machines; but that does not prevent the scheduler from creating a hypothetical schedule that uses 6 machines; that is, the hard constraint is relaxed to do a "what if" analysis that evaluates the benefit of adding a machine to the shop.

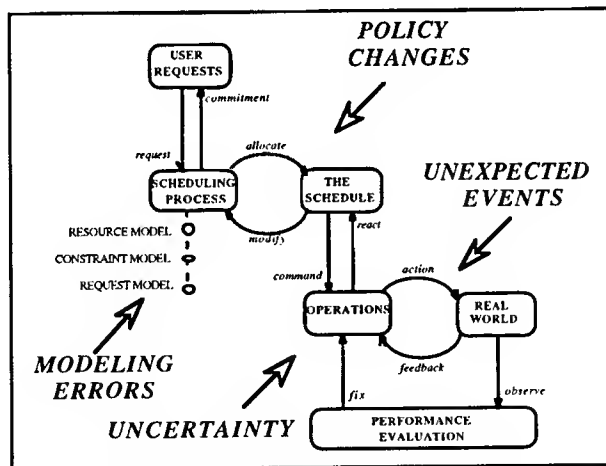
The distinction between hard and soft constraints, and the means by which constraints are relaxed, plays a big role in most scheduling processes. For example, constraints might have *relaxation levels*: when it is discovered that a feasible schedule can not be found all constraints of *level 1* are relaxed (made less strict) and the search for a feasible schedule resumed; if that fails all the constraints of *level 2* are relaxed, etc. Finally, we note that there are other ways to classify constraints<sup>3</sup>, such as discrete vs. continuous, and deterministic vs. probabilistic.

<sup>3</sup> An example just came up: an author called all the system constraints the *implicit* constraints and all the customer preferences the *explicit* constraints.



**Figure B:** The criteria and constraints specify the quality of the scheduling options.

Computer algorithms work with *representations* of the domain, or in other words, with *computer models*. These models attempt to capture the critical features of the criteria, resources, and constraints, so that the algorithms can efficiently evaluate the scheduling options. Intuitively, the hard constraints define the *boundaries* of the search space while the soft constraints define the *shape* of a surface that stretches between the boundaries (i.e.: the hills and valleys of the objective function). The sketch in the figure is a simplification in several ways: 1. dimensionality of a realistic scheduling problem is much higher; 2. there are many non-linear interactions that can occur between suitability functions; and 3. the sketch overlooks the fact that suitability functions might relate to complex interactions among several constraints (e.g.: set up time, integrated maintenance schedule, etc.). But the sketch captures the basic idea: a complicated surface that will not be easy to navigate.



**Figure C:** Many dynamic factors come into play in a realistic scheduling problem.

Computer models are almost always incomplete because of domain *complexity* and limits to what can be effectively represented in data structures. A major contributor to complexity is *uncertainty*. An unexpected event can render a well thought out schedule useless, and at the extreme, if there are lots of unexpected events then scheduling becomes a futile activity. For example, in a factory where machines are constantly breaking down there is really no point to spending vast amounts of time refining the schedule. The assignment of resources would probably be based on a simple rule, such as: put the most important job on the first available machine (until it breaks down). Equipment failures are just one category of unexpected events, others include changes in organizational policy, changes in customer priorities, and changes in due dates. Schedules made according to different policies might vary greatly, and if policy changes are frequent and far reaching, again, the effectiveness of detailed scheduling is significantly reduced.

There are domains where the degree of uncertainty is very small, but even then there is still the issue of modeling a resource accurately, which might require large data structures or a large number of complex

differential equations. The *fidelity* of the models may never reach the level of accuracy that is required, or the computational efficiency may be poor.

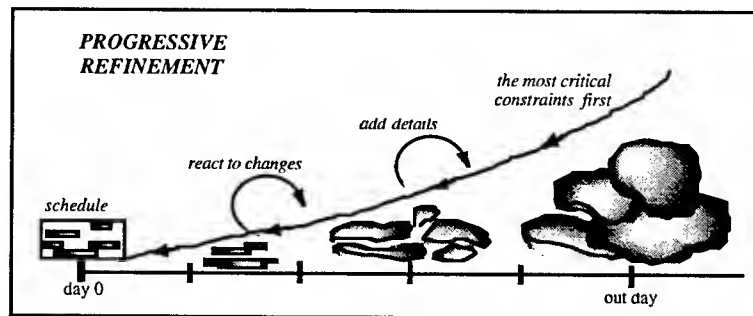


Figure D: Progressive refinement adds the details as the day of operation approaches.

To get a handle on the problems associated with changing criteria, modeling uncertainty and the huge number of scheduling options, certain scheduling strategies are generally recommended, especially *progressive refinement*. Progressive refinement begins with a rough *out day* schedule, and progressively refines it until it is a valid schedule at *day 0* (i.e.: operations). This opens up a variety of hierarchical possibilities. The further the out day the more crude the "schedule" can be, that is, all the hard constraints do not have to be satisfied; for example, there may be some *over booking*, an unspecified communication link, etc. As the *near day* approaches, details can be added and adjustments made, while simultaneously accounting for unexpected events (newly arriving high priority orders, equipment failure, etc.). The schedule may be modified right up to the time of operation. This is also typical of how business trips, and vacations are planned. Three principles related to progressive refinement emerge:

- **Priority Aging:** in a *distributed* environment, when a task is scheduled a *commitment* is made that can spawn many contingent plans, possibly reaching out to other organizations beyond the domain of the current schedule. If the task is suddenly bumped, or moved, the effect can propagate in unexpected and undesirable ways. Intuitively, the task *grows roots* as it sits on the schedule waiting for execution. It is therefore wise to include a factor that raises the *effective priority* of an incumbent task the longer it sits on the schedule, making it more and more difficult to move or bump.
- **Stability:** this concept is related to priority aging, but refers to the degree of task movement we are willing to tolerate to incorporate a new task, or to account for a machine failure, etc. It may be possible to get a new task on the schedule without *bumping* (i.e.: *off* the schedule) any task, but several tasks may have to be modified. It is not wise to be continually shuffling several tasks around (i.e.: avoid *nervous* scheduling) since, for example, a subtle system safety factor might be overlooked, and the humans who interact with the system might become thoroughly confused. Therefore a limit should be put on how many tasks will be moved at any one time.
- **Opportunistic Planning:** unexpected changes usually have an adverse effect on the quality of a schedule, but in some situations a change creates previously unconsidered opportunities. Be on the look out for such *free* opportunities, and capitalize on them.

The results of a successful modeling phase is a recognition of the degree of unavoidable uncertainty, and an identification of the critical parameters that can be used to represent the resources efficiently. If the objective function were well defined, and the models were accurate<sup>4</sup>, the algorithms would be able to explore a well defined search space of scheduling options. The classical literature of *Operations Research* (OR) thrives on such cases, but the number of constraints and options can be so huge that the number of variables exceeds practical limits.

<sup>4</sup> That is, if a *miracle* occurs.

An *algorithm* has to be pretty *smart* to avoid exploring many bad options. Effective *rules* or *heuristics* usually incorporate specific details of the particular domain and consequently such rules do not apply to other domains. This is not a simple matter and it hits at the heart of all heuristic search strategies. We will discuss such heuristics in the following sections.



## 2. Window-Constrained Packing Problem

After investigating several factory scheduling problems we identified a particular packing problem that typifies the difficulties that arise. We call it the "window-constrained packing problem" (WCP). This problem captures the "essence" of many scheduling trade-offs without unnecessary complexity. Many of the results described in this report were first confirmed on this problem and then extrapolations were made to draw conclusions concerning more complex problems.

Many scheduling problems boil down to placing, or fitting, activities onto time lines<sup>5</sup>. For simplicity assume that there is one resource and that the time span involved is the interval  $[0, T]$ . In this ideal model the time line is divided into discrete steps and an activity can stop at the same time as the start of another activity.

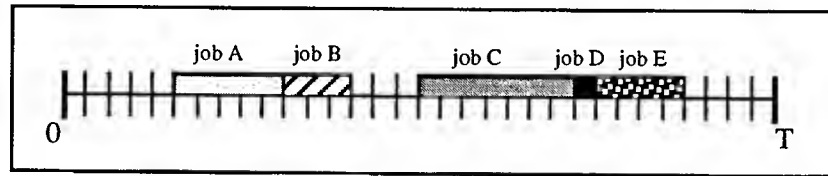


Figure 1: Discrete time steps; a job can start exactly where another job stops.

There are  $n$  jobs  $\{job_i \mid i = 1, \dots, n\}$  and a common constraint is that  $job_i$  must be done within its *window of opportunity*:  $[w_{oi}, w_{fi}]$ . A job is scheduled by assigning it a continuous<sup>6</sup> duration ( $d_i$ ) that satisfies:  $d_{min_i} \leq d_i \leq d_{max_i}$ , where  $d_{max_i}$  and  $d_{min_i}$  are given for each job. That is, each job has a minimum and maximum duration. The scheduled time is referred to as an *activity*.

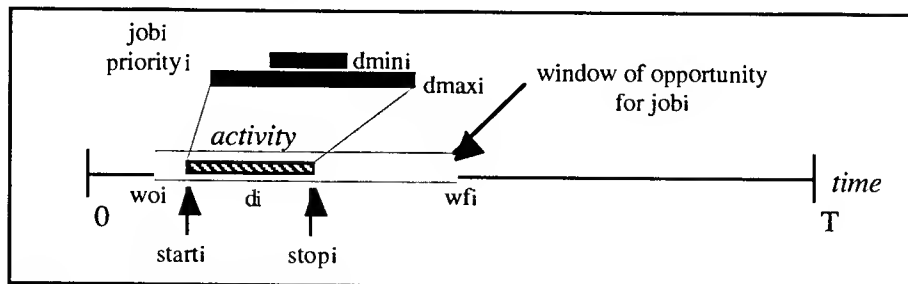


Figure 2: Each job has a priority and must be scheduled (i.e.: become an activity) within its window and duration constraints.

A typical complication is that there can be *preferences* for placements within the window of opportunity. A typical assumption is that the middle of the window is preferred over the edges, so schedules that place jobs

<sup>5</sup> Classical *Knapsack Packing* and *Job Sequencing* problems fall into this category [14].

<sup>6</sup> This is known as the "nonpreemptive" assumption.

in the middle of their windows would be given higher scores. This preference might be specified by a *suitability function*.<sup>7</sup>

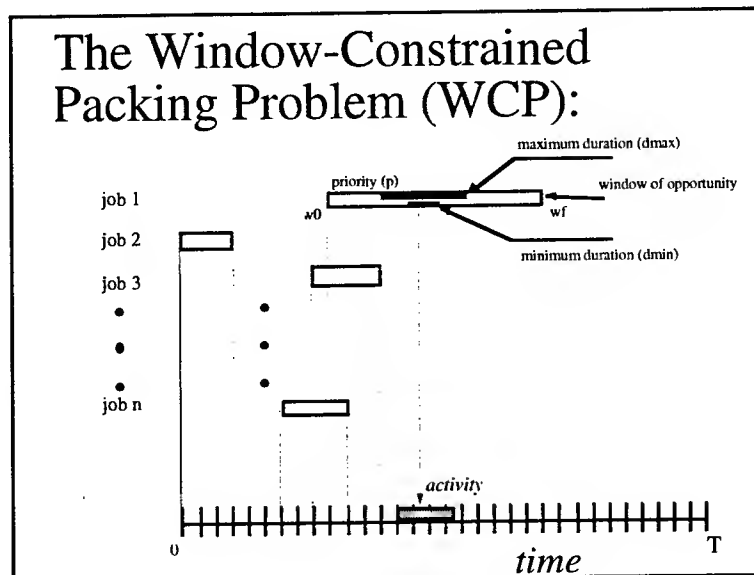


Figure 3: Several jobs compete for time on the resource within their windows of opportunity.

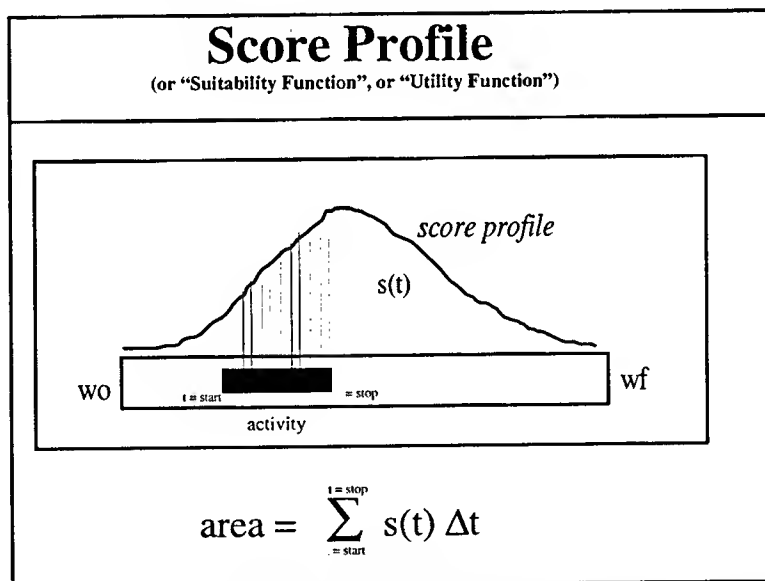


Figure 4: A job is defined by a minimum ( $d_{\min}$ ) and a maximum ( $d_{\max}$ ) duration that must be scheduled within a window of opportunity. The score profile expresses the preference for placement within the window (for the bell curve in this sketch the preference favors the center of the window).

Each job also has a priority assigned to it:  $1 \leq p_i \leq 10$ . The priority measures the relative worth of the job, and might be measured in profit-dollars or other measures of value<sup>8</sup>.

When there is competition for the resource the *scheduler* must determine which jobs get on the schedule and specify the start and stop times that satisfy the window limits and duration constraints.

<sup>7</sup> Sometimes called a *score profile*, or *utility function*.

<sup>8</sup> An example that jumps to mind is the "scientific worth" of a satellite observation activity.

Putting this all together we can formally define the window-constrained packing problem (see figure 5).

**Window-Constrained Packing Problem (WCP):** Given  $n$  objects (requests), each having:

- a window of opportunity  $[w_o, w_f]$ ;
- a suitability function with domain  $= [w_o, w_f]: \{s(t) \mid t \in [w_o, w_f]\}$ ;
- a minimum length ( $d_{min}$ ) and a maximum length ( $d_{max}$ ); and
- a priority  $p$ .

Maximize: 
$$Q = \sum_{i=1}^n x_i \cdot p_i \cdot \text{area}_i$$

Where: 
$$\text{area}_i = \sum_{t=\text{start}_i}^{\text{stop}_i} s_i(t)$$

Subject to the constraints:

$$d_{min_i} \leq d_i \leq d_{max_i}$$

$$d_{max_i} \leq w = w_f - w_o$$

$$w_{oi} \leq \text{start}_i \leq w_{fi} - d_i$$

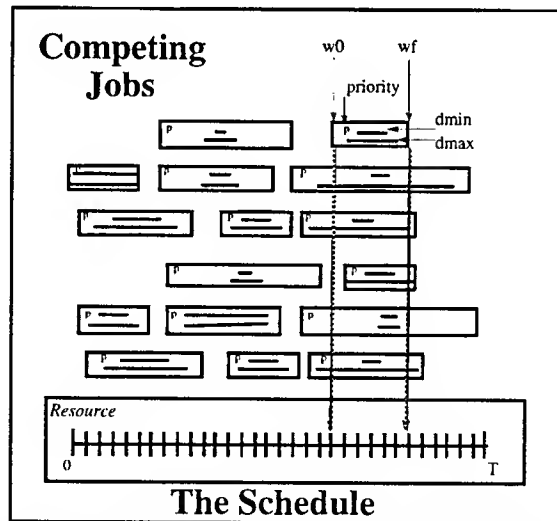
$$x_i \in \{0,1\}$$

(note:  $x_i = 1$  means that job <sub>$i$</sub>  is on the schedule)

$$[\text{start}_i, \text{stop}_i) \cap [\text{start}_j, \text{stop}_j) = \emptyset \quad i \neq j \quad (\text{consistency})$$

**Figure 5:** The details of the Window-Constrained Packing Problem.

The WCP arises as a sub-problem in many scheduling problems (e.g.: job shop, satellites, deliveries, etc.). In the simplest case  $s(t) = 1$  for all the job windows, and  $\text{area}_i = \text{stop}_i - \text{start}_i = d_i$  for all scheduled jobs.



**Figure 6:** The various WCP jobs compete for placement on the schedule.

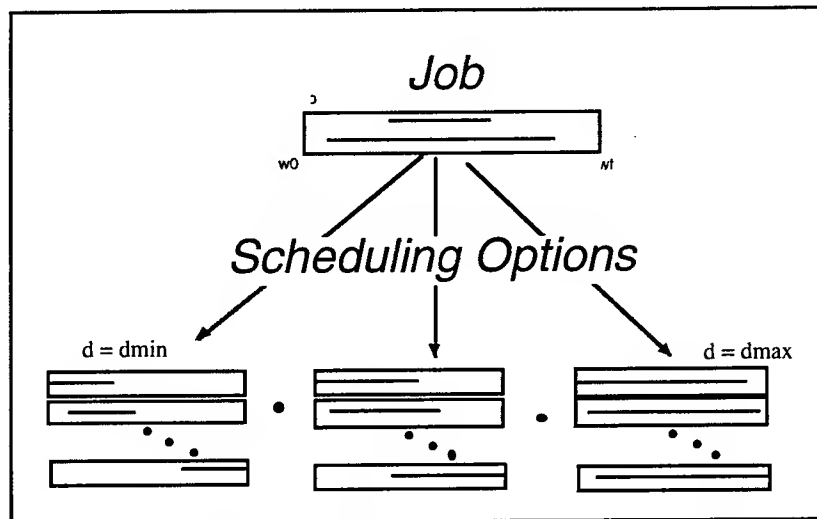


Figure 7: There are usually many scheduling options for each job.

The WCP has the constraints: job duration, window placement, window limits, exclusive use of the resource (i.e.: capacity = 1), non-preemptive activities. A more complicated model would have multiple resources, multiple windows of opportunity, multiple *operations* for a given job, precedence constraints among the jobs/operations, etc. In what follows we stick with the simple WCP, but we have simulated scenarios with the other constraints as well to verify that our "extrapolations" to more complex problems are reasonable. It is safe to say, however, that if we are running into run-time limits on the simple problem (WCP) then we will certainly have worse run-time problems on the more complex problem.

Notice that we are, for the moment, assuming that the only criteria is  $Q$ , which is a function of the individual priorities and associated areas. That is, there is no score related to the other possible criteria, such as:

$$\%Jobs = \frac{\# \text{ jobs scheduled}}{\# \text{ jobs competing}} \cdot 100$$

$$\%Utilization = \frac{\text{time allocated}}{\text{total time}} \cdot 100$$

There are several other criteria that might play a role in evaluating a schedule (e.g.: job spacing), but for now these are good examples of alternatives (or additions) to the  $Q$  defined above. This will become more important later, when we bring up the issue of *multiple* criteria.

In the next sections we introduce specific scheduling *algorithms* and then evaluate their performance on the WCP. We will randomly generate test problems, apply the algorithms and compare the resulting scores.

For very small problem sizes we can run an exhaustive search to determine the optimal solution and use that for comparisons as well. But first we will analyze the # of scheduling options that we might be dealing with (i.e.: the *complexity* of the problem).

**Complexity:** One way to find an *optimal* solution is to exhaustively search the possibilities and compute the scores of all possible schedules. But this can be very time consuming since the average job can have many scheduling options:

number of scheduling options for a job with window width  $w$ , and duration limits  $d_{min}$ ,  $d_{max}$ :

$$\begin{aligned}\# \text{ options} &= 1 + \sum_{q=d_{\min}}^{d_{\max}} [w - q + 1] \\ &= 1 + (d_{\max} - d_{\min} + 1) \cdot (w + 1 - (d_{\min} + d_{\max})/2); \end{aligned}$$

The "1" occurring after the equal sign accounts for the option of *not* scheduling the job. In fact, one of the most critical aspects of this problem is that we do not know ahead of time which jobs should "make" the schedule and which ones to leave off the schedule.

The largest number of options occurs when the window is the largest, the minimum time is the smallest, and the maximum duration is the largest. Let  $w = w_{\max}$ , and suppose  $d_{\min} = 1$  and  $d_{\max} = w$ :

$$(w) \cdot (w + 1 - (w + 1)/2) + 1 = (w) \cdot (w+1)/2 + 1 = (w)^2/2 + w/2 + 1 = O(w^2).$$

So, in the worst case, since there are  $n$  jobs, there can be as many as  $w^{2n}$  total scheduling options.

The simplest situation is when  $d_{\min} = d_{\max} = w$  for each job: 1 option per job. But even this case can get sticky since the jobs may be inconsistent (i.e.: conflict with each other): if the jobs are all consistent there is only 1 total scheduling option; if two jobs overlap then we must choose 1 out of 2, and so on if several overlap.

So, roughly speaking, each job has about  $w$  scheduling options, and there are about  $w^n$  possible schedules. The number of scheduling options grows exponentially as the number of jobs increases (it also grows as  $w$  increases, but at a slower rate). Some of these options will be eliminated because they create inconsistent (non-feasible) schedules. Another extreme is when a job window does not overlap any other job window: the only option that should then be considered for that job is when the duration ( $d_{\max}$ ) is placed to give the maximum area; no other options need be considered since they are guaranteed to produce a worse overall schedule.

Although there are several "special cases" it is clear that the general problem has an exponentially growing number of scheduling options to consider. Finally, notice that the classical Knapsack Problem<sup>9</sup> is the special case:

windows are all the length of the full time interval (i.e.: the "knapsack" =  $[0, T]$ );  
each job had one duration:  $d_{\max} = d_{\min}$ ;  
suitability functions are flat (i.e.: no preferences within a window).

Thus, the WCP is also NP-complete. It is believed (no one has been able to provide a proof) that NP-complete problems do not have polynomial time algorithms that always produce optimal solutions. This forces us to explore necessarily sub-optimal heuristic algorithms. Thus, there is an inherent speed-accuracy trade off.

---

<sup>9</sup> Known to be NP-complete [16].

# 3. Algorithms Overview

---

Here we provide an overview of the many algorithmic approaches that we considered (more details are provided later). The organization of the algorithms into the following categories is not a strict categorization, it just provides a convenient way to refer to them.

## I. Optimal Algorithms:

A. Depth First (DF): Exhaustive search of the options; guaranteed to find the optimal solution; usually takes an impractical amount of time.

B. Branch and Bound (BB): Almost an exhaustive search, but branches that cannot improve on the best result found so far are pruned; usually takes an impractical amount of time.

## II. Constructive Heuristics:

A. Priority Dispatch (PD): Rank the jobs, and schedule them one by one until a complete schedule is constructed; no *backtracking* or *bumping* of scheduled jobs; a very fast but sub-optimal algorithm. Despite the name, we do not have to rank the jobs according to "priority", we can use any features we want (simple to calculate) to rank the jobs.

B. Look Ahead (LA): Proceed as in the Priority Dispatch but when placing a job on the schedule score each of its options by (hypothetically) dispatching the rest of queue (i.e.: apply the PD) and getting a predicted schedule score; it is not optimal, but it does significantly better than the Priority Dispatch; the run time is noticeably longer than the Priority Dispatch but still reasonably fast ( $PD \approx O(n \log n)$ ;  $LA \approx O(n^2)$  ).

C. Fuzzy Logic: We consider fuzzy logic as a means of extending the rule-based aspects of the PD algorithms. That is, we "fuzzify" the rules so that they read:

"when the **PRIORITY** is *high* and the **LAXITY** is *low* then the **RANK** is *high*".

Priority, Laxity and Rank are referred to as "linguistic variables", and they take on "terms" such as *low*, *medium* and *high* [13]. The fuzzy approach adds a layer of abstraction to the rules that is helpful when discussing the nature of the algorithm in a non-technical way. Computations are easily integrated into the approach through specific membership functions and fuzzification/defuzzification steps. Although we have only just begun to incorporate fuzzy logic into the PD, we anticipate that we will gain robustness and flexibility, with little loss of simplicity and speed.

## III. Improvement Algorithms: Start with a schedule (usually the output of one of the constructive approaches) and try to improve it.

A. Hill Climbing (HC): Begin with a schedule and consider ways to change the parameters (move a job, swap two jobs, increase/decrease the duration of a job, etc.). After consider several possible "moves" pick the one that improves the schedule the most, then iterate. Halt when no move can improve the schedule.

B. Simulated Annealing (SA): The same as Hill Climbing but we do not always take the best move, in fact we sometimes (probabilistic decision) take a move that decreases the quality of the schedule. As the iterations progress we bias the decisions in favor of Hill Climbing and ultimately converge to pure Hill Climbing.

C. Tabu Search (TS): As we consider the moves, as in the Hill Climber, we keep a list of moves that are not allowed to be repeated within a certain number of iterations. Like Simulated Annealing we will not halt at local optima, but unlike Simulated Annealing we maintain a *deterministic* search strategy (as opposed to the randomness of SA).

D. Repair: These algorithms are variations on the previous heuristics but the assumption is that we are dealing with a schedule that has been suddenly rendered incorrect by some event (e.g.: machine failure, a canceled order, etc.). This may boil down to making a few simple adjustments to the schedule (e.g.: slide all the jobs on the failed machine forward to the next available machine), or it might involve the re-scheduling of most of the scheduled jobs. In many cases the "stability" of the schedule comes into play. The "stability" measures the number of changes in the schedule over a short period of time. It is usually undesirable to be rapidly changing the schedule (even if the "score" of the new schedule is better in some other sense). A typical strategy would be to only reschedule a few selected jobs, or the only the lowest priority jobs.

There are a couple of other terms that come up in the literature that are something like "repair": "incremental" and "reactive". The incremental approach is similar to the repair approach, except that it implies the day-to-day scheduling process: adjusting jobs to meet the constraints, fixing an oversubscription problem here, a missing resource allocation there, etc. We bring up such issues under the title "progressive refinement" later in this report. Reactive scheduling implies the immediate response to an "emergency" situation, like a "reflex action", and is thus almost identical to what we call repair-based, only differing in the amount of time and thought required to make the repair.

IV. Genetic Algorithms: A "population" of schedules is allowed to evolve by breeding new schedules from the better schedules seen so far. We tested two particular GA's:

A. Indirect GA (GA): Each member of the population is a *permutation* of the list of jobs. The members are evaluated by feeding the permutation into either the PD or LA algorithms (or any other constructive method). Feeding the permutations into the PD will run through the populations faster than with the LA, but the LA will tend to give better results in fewer generations. We explored a few "crossover" operators and the results presented below are with the PMX operator [17]. In this representation a member of the population is a permutation, and the permutation is used to construct a schedule (as opposed to using the schedules themselves as members of the population) so it is referred to as an "indirect" representation. More details (mutations, etc.) are provided later in this report.

B. Augmented GA: We ~~added~~ more information to the Indirect GA. Each permutation is augmented with 0's and 1's to indicate which jobs are allowed to be scheduled. That is, a 1 indicates that the dispatcher should consider scheduling the job, and a 0 that it should not. We augment the crossover to pass these 0's and 1's along with the generations. This modification improves the GA results significantly.

After running many of versions of these algorithms on the WCP it became clear that the WCP has three critical pieces:

- Which jobs make the schedule?
- What *order* should they be in?
- What is the *exact start and stop time* for each scheduled job?

These questions are not independent, but can be interpreted in a hierarchical way. First of all, if we knew ahead of time *which* of the contending jobs were going to make the schedule (a subset of the total set of competing jobs) then finding the optimal solution to the WCP problem would be significantly easier. Given the jobs that are to make the schedule, then if we knew the *ordering* of those jobs on the resource the problem we be even easier (the only thing then left to do is to slip/slide/grow/shrink the jobs until we get the highest score we can get<sup>10</sup>).

It is interesting to see how each of our algorithmic methods deals with these questions. For example, the PD answers the first question, implicitly, by ranking the jobs, then answers the second question with its allocation rules, and finally answers the third question by "growing" the jobs that made the schedule. It is more difficult to see these separate issues in some of the other algorithms but we have found that, in general, these three questions help simplify the analysis.

Even though we have this modular hierarchy of concerns, we must keep in mind that a given set of jobs can compete for a resource in a possibly large number of ways. Thus, unless we check *all* interactions between jobs and evaluate *all* potential conflicts we cannot be certain that we have the optimal solution. Such is the nature of NP-Complete problems. Another way of saying this is to point out that for any proposed heuristic a set of jobs can be created that will force the heuristic to miss the optimal solution (possibly by a huge amount). Therefore, the effectiveness of a heuristic depends on the nature of the particular set of competing jobs.

---

<sup>10</sup> We are assuming that the suitability functions are simple (e.g.: unimodal) and that a simple Hill Climbing approach would be very effective; if the functions are complex then this step is much more difficult.



# 4. Optimal Algorithms

## Depth First Branch and Bound

The "optimal" algorithms are guaranteed to produce optimal solutions (according to a well defined criteria), although they may take a long time to do so. Most optimal algorithms are impractical in real applications but they often serve as theoretical tools for grasping the complexity of the problem. We will discuss two such algorithms<sup>11</sup>.

**Depth First (DF):** A *Depth-First* search (see figure 8) would be one way to exhaustively search all possible schedules and find the highest scoring one (the optimal schedule). But such a search would be unnecessarily time consuming in most cases. The Branch and Bound algorithm described below uses knowledge of the best schedule so far to prune away fruitless branches of the search.

**Branch and Bound (BB):** One way that we might eliminate branches is by somehow knowing that a branch cannot possibly give us a result better than one we already have (see figures 12a and 12b).

Suppose we knew that there is at least one schedule with a score of  $Q = x$ . If we are constructing a schedule, job by job, we can use  $x$  to detect bad branches: when considering an option, compute the total score that would be gotten if all the *remaining* requests were scheduled at their optimal placement (maximum duration, window middle, without checking for conflicts with other unscheduled jobs). If this value is less than  $x$  we can safely prune that branch off the tree of choices.

To be more specific: given that the jobs are indexed as usual  $\{\text{job}_i \mid i = 1, \dots, n\}$  and we are sequencing in a depth first sense through the job options in that order, let:

$$Q_{k-1} = \sum_{i=1}^{k-1} x_i \cdot \text{score}_i$$

where  $\text{score}_i = p_i \cdot \text{area}_i$

$$\text{and } x_i = \begin{cases} 1 & \text{job}_i \text{ is on the schedule} \\ 0 & \text{job}_i \text{ is not on the schedule} \end{cases}$$

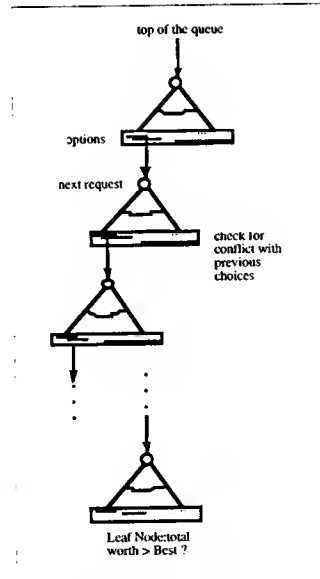
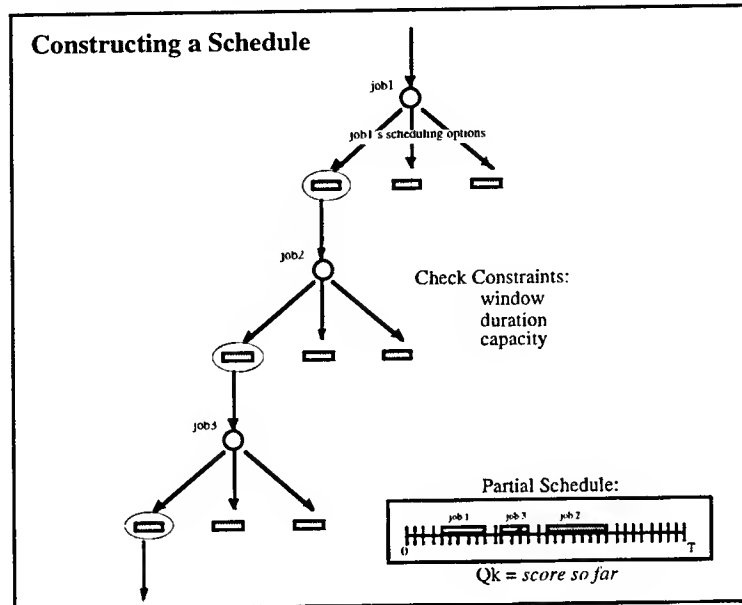
When we consider  $\text{job}_k$  we can score its possible placements by computing:

$$\text{placement score} = q_k = Q_{k-1} + \text{score}_k + Q_{\text{maxk}+1}$$
$$Q_{\text{maxk}+1} = \sum_{j=k+1}^n p_j \cdot \text{maxarea}_j$$

$\text{maxarea}_j$  = the largest possible area achievable by  $\text{job}_j$   
(usually the maximum duration placed in the window middle).

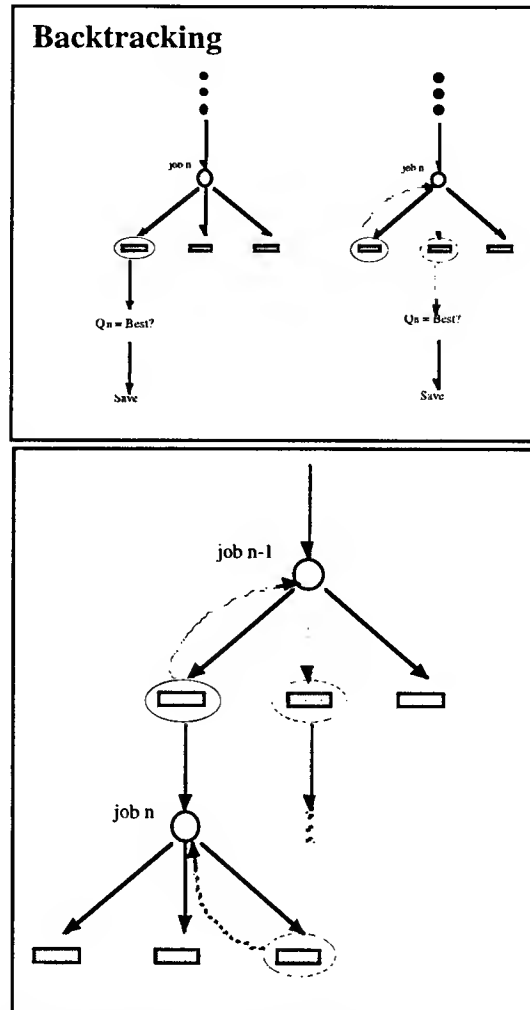
<sup>11</sup> These algorithms might be called "backtracking" algorithms.

$Q_{k-1}$  is the score of the partial schedule up till the  $k^{\text{th}}$  job is considered,  $\text{score}_k$  is the score of a particular placement of  $\text{job}_k$ , and  $Q_{\text{maxk}+1}$  is the maximum possible score of all the jobs later in the queue:  $\text{job}_j, j > k$  (without considering conflicts). Notice that as we score the options  $Q_{k-1}$  and  $Q_{\text{maxk}+1}$  do not change. We only consider options that do not conflict with the previous scheduling decisions. If the score of an option ( $q_k$ ) is less than the score of the best schedule so far then we prune that option and its branch.

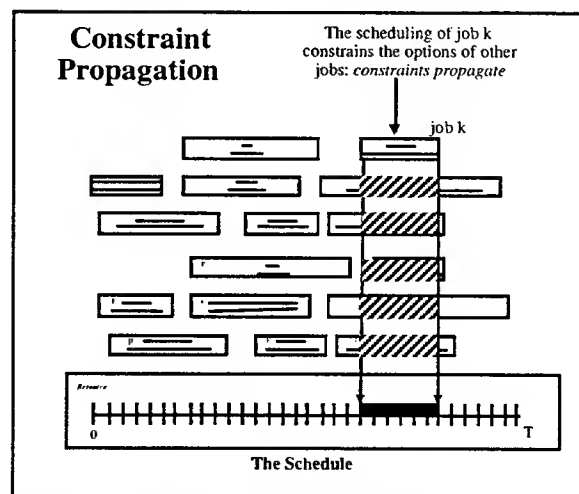


**Figure 8: Depth First Search.** Sort the requests. Consider all scheduling options by taking the top of the queue, finding a scheduling option, then go on to the next request. When a leaf node is reached: evaluate the schedule (score) and see if it is the "best so far". Then backtrack to other possibilities<sup>12</sup>.

<sup>12</sup> Remember to consider the option of not scheduling the job.



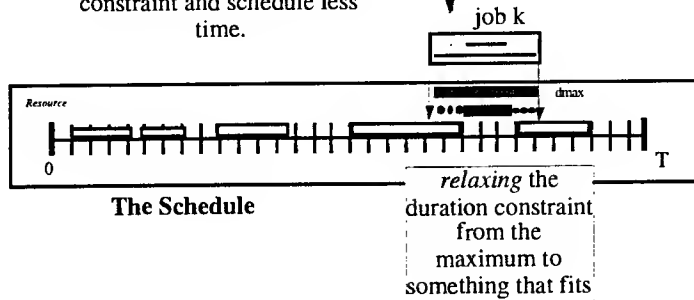
**Figure 9:** Backtracking in a depth first search begins when we hit the last job. Then we try all the options for the last job, and when they are depleted the search backtracks to the previous job and considers its next option, and so on.



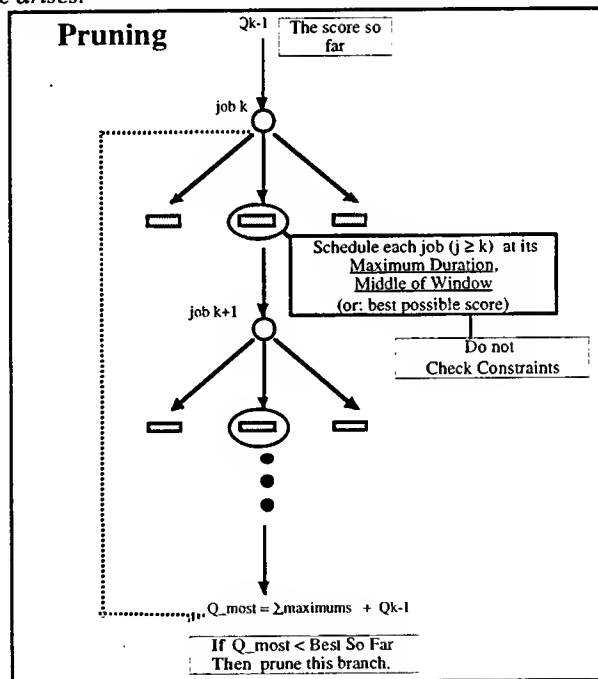
**Figure 10:** Constraints propagate as scheduling decisions are made. Some jobs are not affected, and some have no options left.

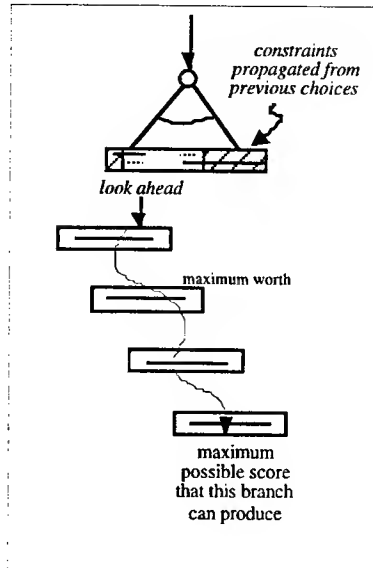
## Constraint Relaxation

Want to schedule  $d_{max}$ , but it won't fit. So *relax* the duration constraint and schedule less time.

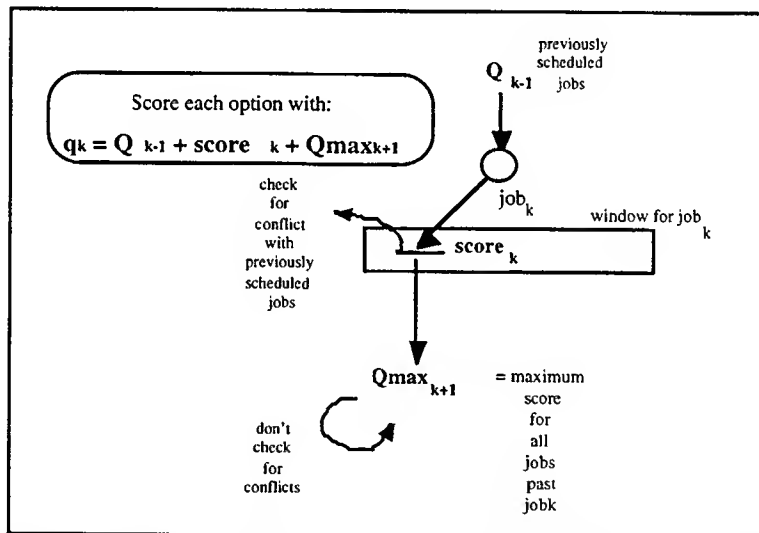


**Figure 11:** A simple example of "constraint relaxation". In this case job  $k$  wanted to be scheduled for the maximum duration, but the previous scheduling commitments would conflict, so the duration is relaxed to a length that fits the available time. In general, constraints are relaxed from higher levels of satisfaction to lesser ones as the need arises.



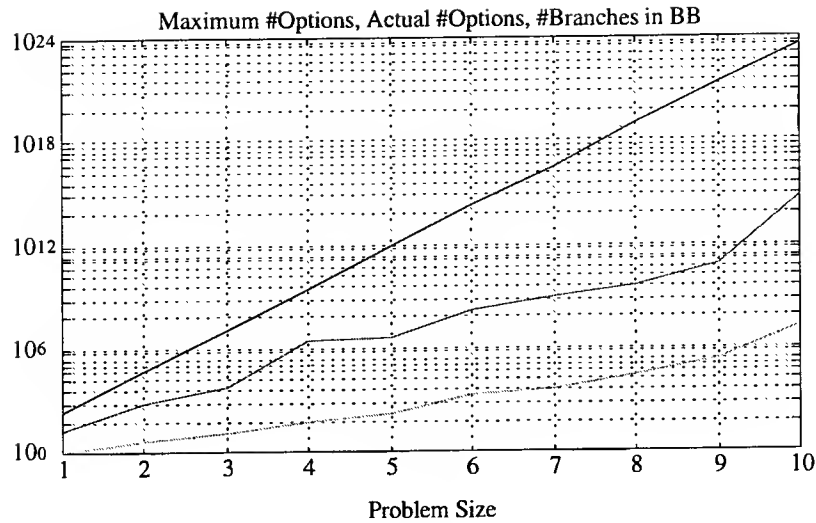


**Figure 12a:** A way to prune some branches in the BB algorithm. At each scheduling option assume that all the remaining requests are scheduled at maximum duration and at there highest scoring positions within their windows of opportunity without regard for conflicts. If the resulting score is less than the highest scoring scheduling saved so far then there is no point in pursuing this branch

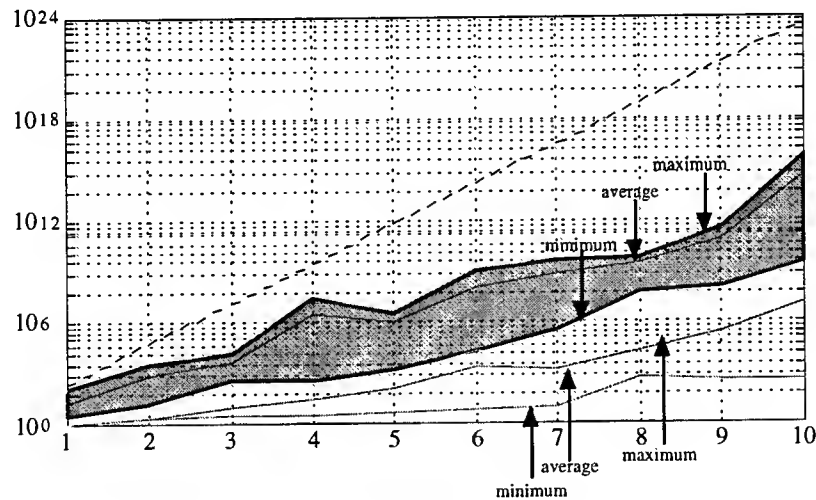


**Figure 12b:** Branch and Bound: score each scheduling option by adding the score of the partial schedule and the score of the option and the maximum possible score of the remaining jobs. If the score of an option is less than the best schedule so far prune that branch of the search tree.

Branch and Bound might still have to search all the possibilities, it depends on the particular jobs. However, pruning away "bad" options (branches of the search) can save time, without compromising the guarantee of optimality. The BB algorithm will take a relatively long time to run and will be restricted to problems of size  $n \leq 10$  (feel free to push back the wall of exponential growth). The requests can be sorted at the start to make it easier to prune paths. Estimated order:  $O(w^n)$ .



**Figure 13a:** The case considered here is  $w_{\max} = 25$ ,  $w_{\min} = 5$ ,  $T = 200$ . The top curve is the worst case number of scheduling options for each problem size. The middle curve is the actual number of scheduling options for randomly generated job sets (average of 10 runs per problem size). The bottom curve is the number of scheduling options explored by the BB, i.e.: branches of the BB algorithm. Although the BB cuts the number of branches explored by a large amount, the exponential growth of the BB algorithm is evident. When  $N=10$  the run times for the BB can be in the hours.



**Figure 13b:** A follow up on the previous figure, showing the worst case numbers for the 10 runs per problem size.

# 5. *Constructive Heuristics*

## *Priority Dispatch Look Ahead*

---

Constructive heuristics *build* a schedule from *scratch*. They usually use rules for *selecting* jobs and rules for *allocating* resources to the jobs. They can also be used to "add" to an existing schedule, usually treating the existing commitments as hard constraints. The first constructive heuristic that we describe is called the Priority Dispatch method, a method with many possible variations. The main advantage of this method is that it is very easy to understand (three distinct *phases*) and runs very fast. The Look Ahead is similar to the Priority Dispatch approach but uses a much more "intelligent" allocation step. The Look Ahead approach tends to uncover the interactions between jobs in the queue and therefore anticipates certain conflicts.

**Priority Dispatch (PD):** This method begins with a batch of unscheduled jobs and uses three phases: *Selection*, *Allocation*, and *Optimization*, to construct a schedule. There are many possible variations on these steps, and here we present a representative version.

**Phase I: Selection.** Rank the contending, unscheduled, jobs according to:

$$\text{rank}(\text{job}_i) = f_i, \text{ where: } f_i = \frac{p_i}{\text{dimin}}$$

This, of course, is just one way to rank the jobs. Experience shows that this is a very good way to rank the jobs (for randomly generated sets). Intuition explains this by the fact that the feature  $f$  measures the "worth per unit length" of the job<sup>13</sup>, so jobs with a high value of  $f$  tend to have a high value of  $p$  mixed with a small duration. Such jobs make relatively large contributions to the overall score while conflicting with very few other jobs (remember, we are speaking *intuitively* here).

**Phase II: Allocation.** Take the job at the top of the queue from phase 1 and *consider* scheduling the minimum duration ( $d_{\min}$ ). If there is room for it within its window of opportunity (that does not conflict with previously scheduled jobs) consider the *best* spot: where the area under the score profile is largest. If there is no room for the job skip to the next job in the queue<sup>14</sup>. By choosing to place  $d_{\min}$ , as opposed to longer durations (such as  $d_{\max}$ ), we are getting the job on the schedule while conflicting with the least number of other jobs<sup>15</sup>. This is an interesting balance of "greedy" and "altruistic" strategies:

---

<sup>13</sup> This is closely related to the "weighted shortest processing time (WSPT)" algorithm that is proven optimal for specific single machine sequencing problems [14].

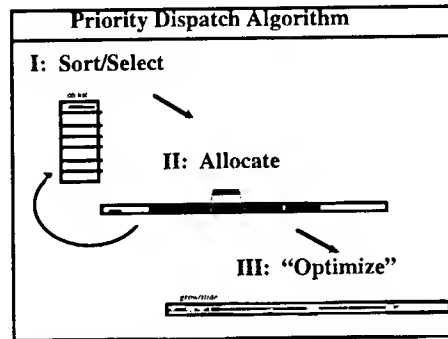
<sup>14</sup> We do not consider *bumping* a scheduled task here. Such a variation can be easily added, but we have found this to be effective only when the job queue is poorly sorted (i.e.: using a bad sort key).

<sup>15</sup> We know from prior simulations that this favors getting *more jobs* on the schedule at the expense of some *worth*. Getting more jobs on the schedule is not always mentioned as a criteria but it almost always is beneficial.

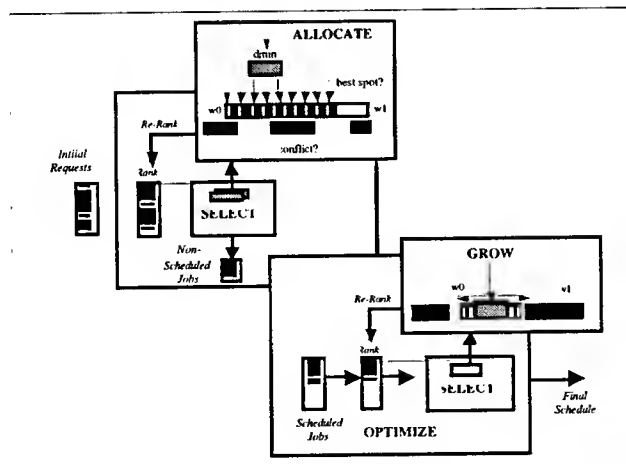
place the job at its *best* spot but the job only gets its *minimum* duration. Continue through the job queue<sup>16</sup> until all jobs have had one chance to be scheduled.

**Phase III: Optimization.** After phases I and II are complete, a final phase runs through the scheduled jobs, this time ranked according to  $\text{rank}(\text{job}_i) = p_i$ , and *grows* each job to the left and right until it meets another activity, reaches its maximum allowed duration ( $d_{\max}^i$ ), or reaches the edge(s) of its window of opportunity. This is a form of Hill Climbing, and in fact, any hill climber algorithm could be plugged in at this step, but the theme of the PD is to keep it simple. Estimated order of the PD algorithm:  $O(n \log n)$ .

The advantages of the PD approach are that it is nicely modular, the scheduling strategy is easy to understand, it runs very fast, and produces acceptable schedules most of the time. The modularity and simplicity support the addition of many variations that may prove useful for a specific application. The accuracy of the PD, however, can be very poor for certain sets of competing jobs. In some such cases additional rules can be added to the Selection and Allocation phases. If the desired accuracy still cannot be achieved, we recommend going to the Look Ahead method, which sacrifices some speed and simplicity to get a significant increase in accuracy.



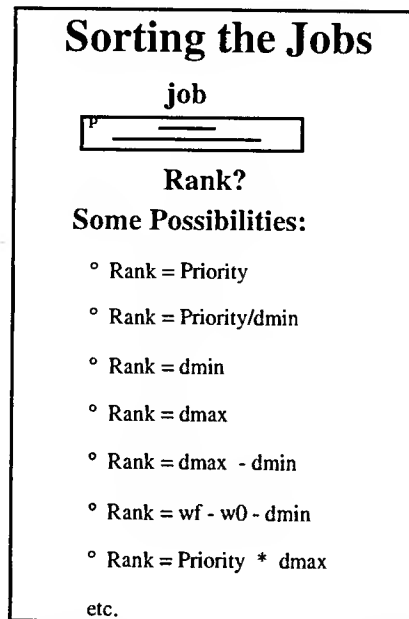
**Figure 14a:** The PD approach uses three phases. The Sort phase ranks the jobs and the Allocate phase places them on the schedule. When all the jobs are allocated the Optimization phase does some fine tuning (e.g.: take advantage of unused resources).



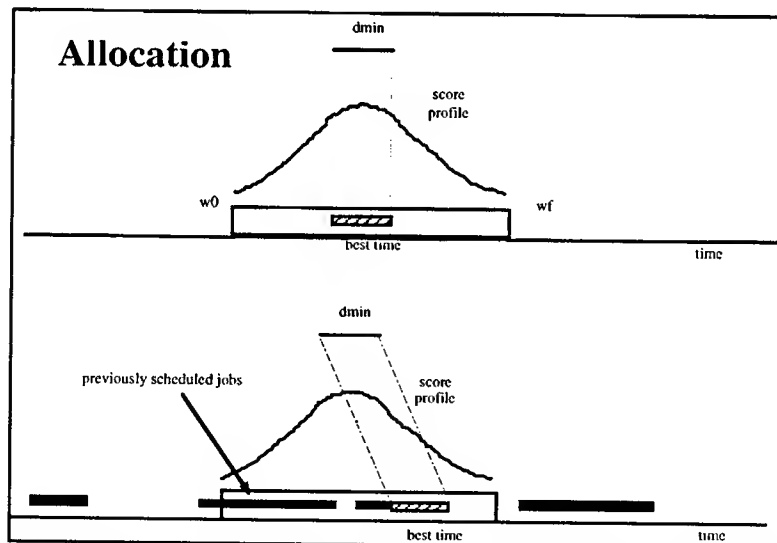
**Figure 14b:** There are many variation on the PD phases. In this sketch the PD ranks the jobs and after each allocation re-ranks the jobs (their ranking features may be affected by a job that just got scheduled). The allocation phase allocates only the minimum duration to the job, and the optimization phase "grows" the jobs that made the schedule into unused spaces.

<sup>16</sup> There are cases where we re-sort the queue after each allocation (e.g.: when the allocation of a job changes the rankings of the jobs remaining in the queue).



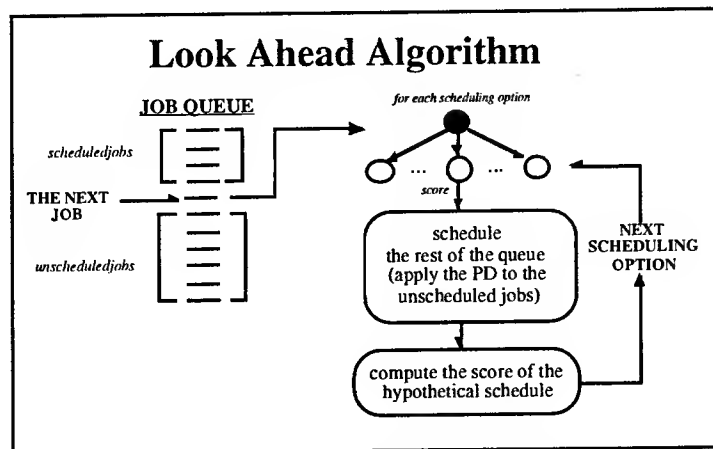


**Figure 15:** There are many ways to sort the jobs. In most cases the parameters of the job are used to compute a value that is used as the sort "key". In some cases the sort key depends on the current "state" of the schedule; for example if  $\text{rank} = [(\text{time available for scheduling}) - d_{\min}]$ , then the value will change as jobs are scheduled that use some of the available time.

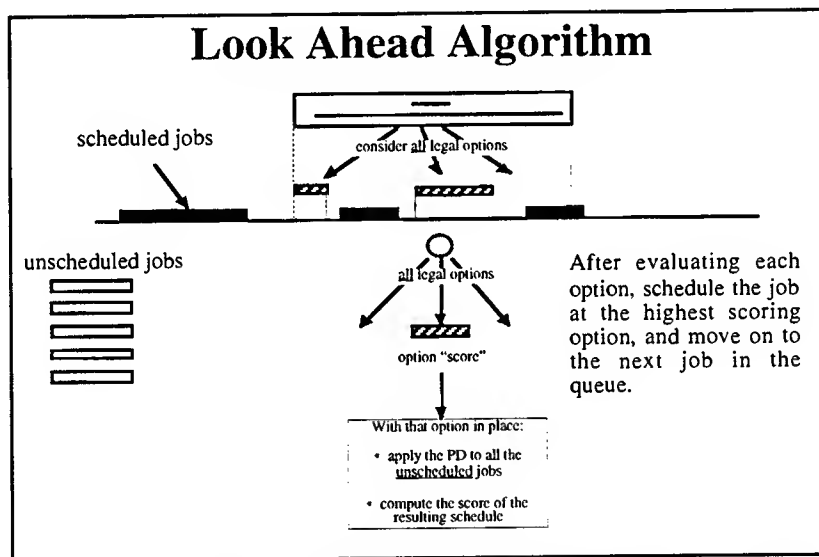


**Figure 16:** The "allocation" step in the PD uses a maximize-minimize strategy: although the minimum duration is scheduled, it is scheduled at the best available time (with respect to its score profile).

**Look Ahead Algorithm (LA):** A powerful modification can be made to the PD method at the allocation step, when the job is placed at the *best* spot (as determined by area under the suitability function). This can be greatly improved by determining *best* by *looking ahead*: To score the possible placements of a job let the dispatch queue continue scheduling (without actually making the allocations, but obeying all the constraints) and compute the final Q. Do this for each possible placement of the job. The position that scores the highest is declared "best" and the job is allocated there, and the iteration continues (no backtracking). This accounts for interactions of the job with as yet unscheduled tasks (*down-stream* jobs). This adds a noticeable amount of processing time but we gain a significant amount of quality. Estimated order:  $O(n^2)$ .



**Figure 17:** The Look Ahead algorithm scores each scheduling option by hypothetically scheduling the rest of the queue (using the PD) and computing the score of the hypothetical schedule. The option that scores the highest is the option that is actually scheduled and then the system moves to the next job in the queue.



**Figure 18:** It is important to realize that the Look Ahead Algorithm considers all the legal scheduling options for each job, and is not restricted to considering the minimum duration as is the PD.

The advantages of the Look Ahead algorithm is that it is an excellent trade off of speed and simplicity for accuracy. The algorithm is relatively fast, modular, relatively easy to understand, and achieves its accuracy by dynamically detecting down-stream conflicts. The effectiveness of this algorithm highlights the fact that the interactions (i.e.: competition) between the contending jobs must be evaluated in detail to uncover the dependencies. Intuitively, the LA is measuring the sensitivity of the down-stream jobs to the particular placements of the job currently being allocated.

## 6. Fuzzy Logic

We treat Fuzzy Logic as an extension of the rule-based approach. For example, consider the following rule:

If **Priority** is *high* and the **Laxity** is *low* Then the **Rank** is *high*.

From a "fuzzy" point of view Priority, Laxity and Rank are *linguistic variables* that are qualified by *terms* such as: {*low, medium, high*}. Membership in the qualifying group is measured by a *membership function* such as those shown in figure 19.

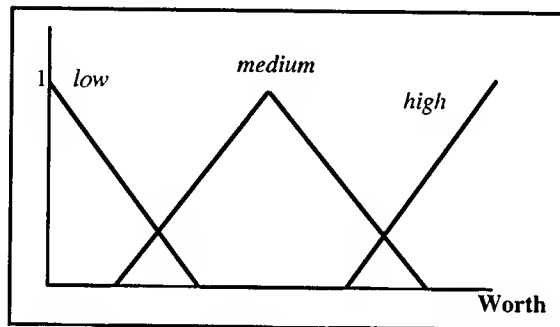


Figure 19: A typical set of fuzzy membership functions for the linguistic variable Worth.

For a given set of jobs we can compute specific values for Priority and Laxity with formulas such as:

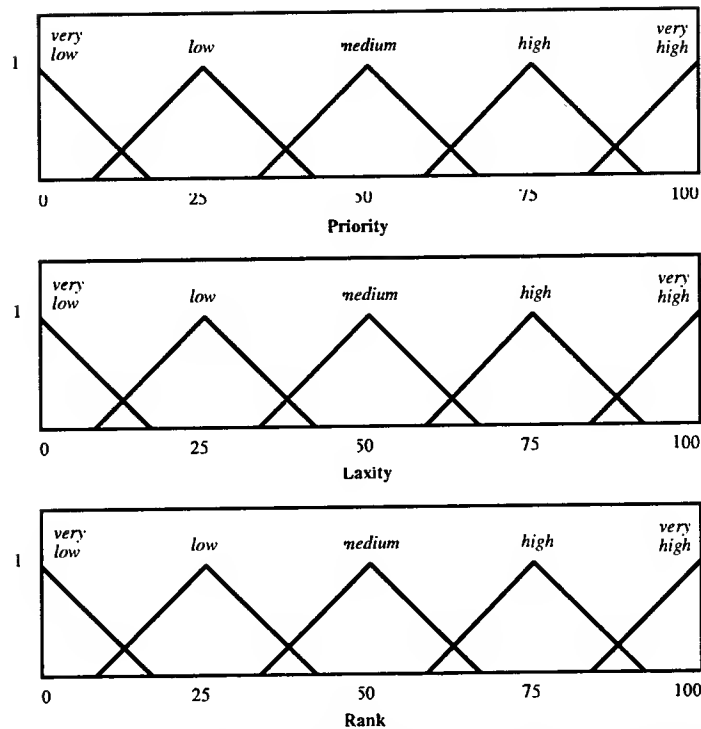
$$\begin{aligned} \text{Priority}(\text{job}_i) &= p_i \cdot d_{\text{imax}} \\ \text{Laxity}(\text{job}_i) &= W_i - d_{\text{imin}} \end{aligned}$$

Where  $p$  is the job's priority<sup>17</sup> and  $W$  is the width of the job's window of opportunity (WCP). The range of values of Priority and Laxity over a given job set would be used to calibrate the membership curves (this provides the appropriate labeling the x-axis: minimum to maximum values). With these values we can generate rankings for the jobs using standard fuzzy logic: "fuzzy and" and "fuzzy or" [13].

Rules of Thumb
<b>Top of the Queue:</b> High Priority Low Laxity Short Jobs
<b>Bottom of the Queue:</b> Low Priority High Laxity Long Jobs

Figure 20: Some basic rules of thumb that lend themselves to a fuzzy implementation.

<sup>17</sup> We are "overloading" the term "priority". In the fuzzy sense (Priority) we think of it as a generic term for the value of the job. In the original sense "priority" is the specific priority assigned to the job when it is accepted into the job queue.

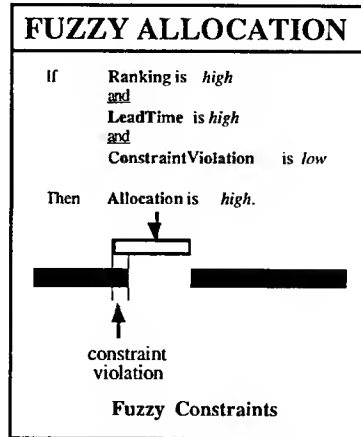


**Figure 21a:** By scaling the measured values of *Priority*, *Laxity* and *Rank* to the range [0,100], we can compute values (crisp) and fuzzify them into the range [0,1].

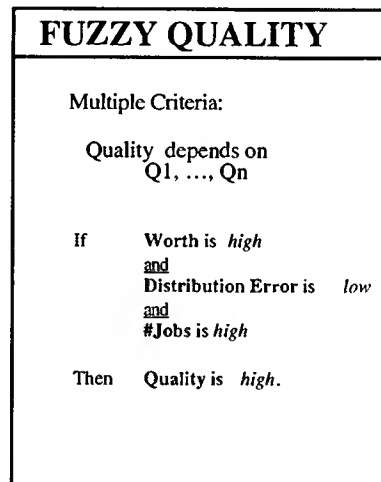
	Priority		Laxity		Rank	
	VL	L	M	H	VH	
VL	VL	L	M	H	VH	
L	M	M	L	L	VL	
M	H	H	M	M	L	
H	VH	VH	H	M	M	
VH	VH	VH	VH	H	H	

**Figure 21b:** If we introduce the linguistic terms: very low, low, medium, high, and very high, we can create 25 fuzzy rules for the various values of *Priority* and *Laxity*. For example, the top left entry in the chart says: "if the *Priority* is very low (VL) and the *Laxity* is very low (VL) then the *Rank* is low (L)".

The chart in figure 21b expresses the strategy that the ranking of a job should be lowered if the laxity is high and raised if the laxity is low. Looking at the first column in the figure 21b we see that the laxity is very low and the priority ranges from very low to very high, while entries in the chart indicate that the ranking is one step above the priority: when the priority is very low the ranking is low, when the priority is low the ranking is medium, etc. A similar thought is expressed by the other columns in the chart.



**Figure 22a:** We can go further with the fuzzy concept and apply it to the "allocation" step in the dispatch approach. As the figure indicates, the constraints themselves can be modeled in a fuzzy way, allowing degrees of constraint violation depending on such things as the lead time and importance of the job being allocated.



**Figure 22b:** In situations with multiple criteria for measuring the quality of a schedule, we can apply fuzzy concepts for combining the disparate measures of quality.

When we put the fuzzy ranking, fuzzy allocation, and fuzzy quality together we get what might be called "fuzzy scheduling". The idea of "fuzzy scheduling" fits nicely with the progressive refinement approach described later in this report: the constraints can become less and less fuzzy as the day of operation approaches.

The *fuzzy* approach has two attractive features:

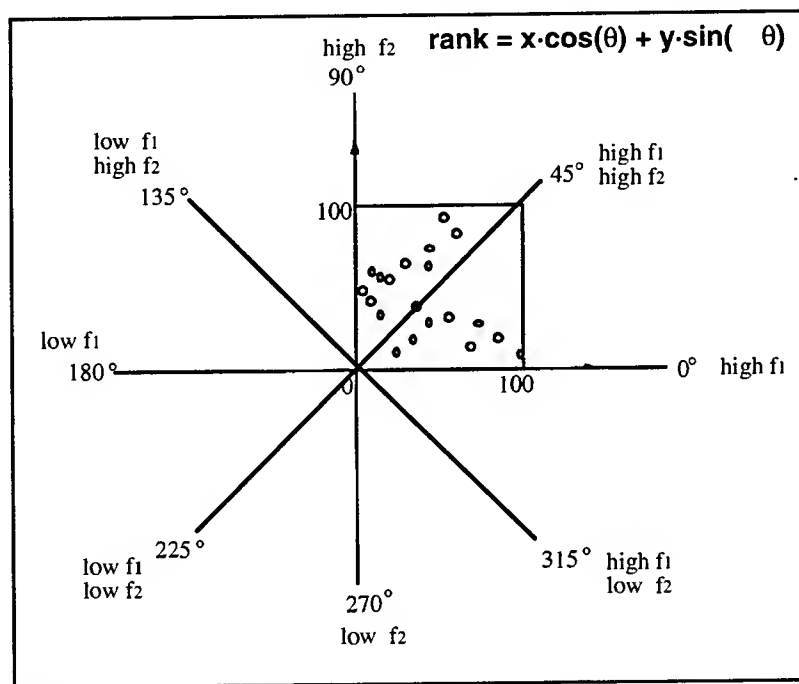
- It adds a "layer of abstraction" that allows the designer to discuss the nature of the rules in a simple way.
- It has a built in tendency to be insensitive to small disturbances (the "fuzzy and" and "fuzzy or" operations tend to screen out small changes).

The disadvantage to the fuzzy approach are:

- The membership functions are not always easy to define and calibrate.
- It requires some additional processing time.

We believe that the advantages will outweigh the disadvantages when fuzzy logic is applied to the PD rule structure. Intuitively, the fuzzy approach seems consistent with the fact that the objective function (e.g.: a combination of the measures of all the scheduling criteria) is not as accurate as we would like it to be. It tends to have subjective weightings of parameters that is more like adding *apples* and *oranges* than a precise mathematical representation. Add in the fact that the criteria can be vague and can be rapidly changing, and it appears that a fuzzy approach would help filter out meaningless variations. We are currently investigating this approach.

Finally, reference [6] introduces the " $\theta$ -projection". This method serves two purposes: 1. it is an initial version of a fuzzy logic approach to the PD; and 2. it emphasizes the "dynamic" aspect of the PD: the queue can be sorted after each scheduling decision without a significant amount of additional processing time.



**Figure 23:** Consider the case of two "features",  $f_1$  and  $f_2$ , used to rank the jobs (e.g.: worth and laxity). After the features are normalized to the  $[0,100]$  range each job is represented by a point in the square in the figure. If the points are projected onto the various lines ( $0^\circ$ ,  $45^\circ$ , etc.) we would be sorting the jobs according to a "high" or "low" measure of the features. For example, projecting onto the  $0^\circ$  line is the same as ranking the jobs exclusively according to  $f_1$ , projecting onto  $90^\circ$  sorts them by  $f_2$ , projecting them onto  $45^\circ$  sorts them by  $1/2 \cdot f_1 + 1/2 \cdot f_2$ , etc. Also notice that if we project onto  $270^\circ$  then we are sorting according to  $f_1$  but in ascending order. This is the basis of the  $\theta$ -filter method described in [6].

Additionally, the points in the square in figure 23 would be *moving* if we were re-ranking the jobs after each allocation step. For example, if we were ranking the jobs so that a high  $f_1$  and high  $f_2$  were at the top of the queue (i.e.:  $\theta = 45^\circ$ ) then the unscheduled jobs would be migrating to the upper left of the square.

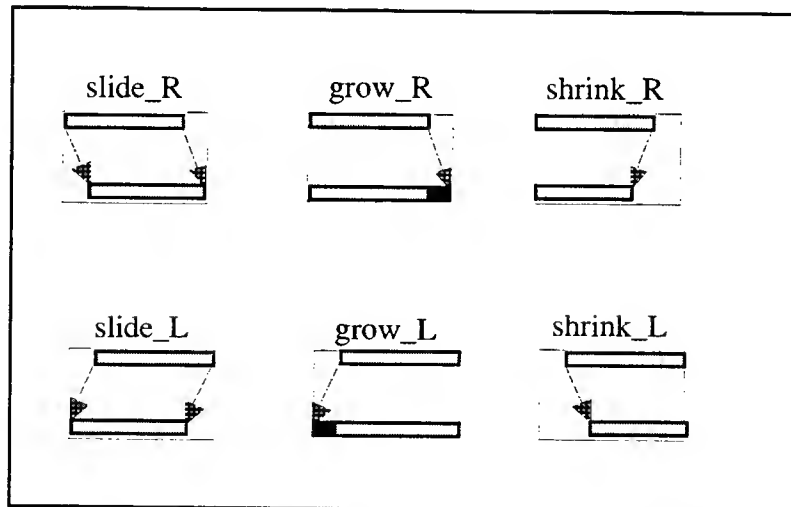
# 7. *Improvement Heuristics*

## *Hill Climbing* *Simulated Annealing* *Tabu Search*

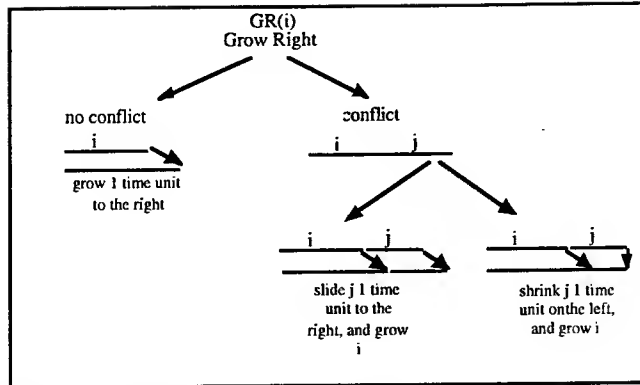
---

There are many algorithms that would be classified as "improvement"-based: begin with a schedule and successively improve it. One way to improve the schedule might be to strip everything off the schedule and then rebuild it with a constructive heuristic, so it is clear that the distinction between "improvement" and "constructive" can be vague at times, and to some extent they represent two ends of a full spectrum of methods. For example, the PD algorithm uses a mix of improvement and constructive techniques: the first two phases are constructive, while the third phase is improvement-based (the "optimization" phase). The optimization phase of the PD is, in fact, a form of Hill Climbing.

**Hill Climbing (HC):** Hill climbing algorithms try to improve the quality of a given schedule by a small change in one or several of the variables (i.e.: *moves*). Such an algorithm could be used in the optimization phase of the PD algorithm. The variables in this case are the start and stop times for each of the jobs on the schedule. For the WCP we know that there may be jobs that are *not* scheduled, so we could consider bumping jobs off the schedule and placing new ones on, but for this analysis we will fix the jobs on the schedule. Neither do we consider *swapping* the order of any jobs. Thus, for now we only consider *sliding*, *shrinking*, and *growing* the jobs by one time unit (within the window, duration, and capacity constraints).



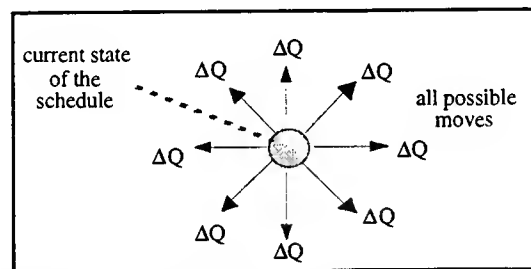
**Figure 24:** Some of the "moves" that might be applied to a job that is on the schedule. Notice that the moves must obey the widow and duration constraints. The Slide and Grow moves might conflict with a neighboring job on the schedule, in which case the decision must be made to either not allow the move or accommodate it by modifying the neighbor (e.g.: slide, shrink or bump the neighbor).



**Figure 25:** Hill Climbing is complicated by the fact that an incremental change in one parameter may cause a ripple effect as it affects other jobs on the schedule. For example, growing a job to the right might conflict with a neighboring job, as shown in the figure (job  $i$  bumps into job  $j$ ).

A change in a variable induces a change in  $Q$  (the measure of schedule quality). In theory we would consider all possible changes (one time unit) and take the best one:

$$\max_v \{ \Delta Q / \Delta v \}$$



**Figure 26:** Hill Climbing considers all possible incremental changes in the state of the schedule and then picks the one that increase  $Q$  by the most (largest  $\Delta Q$ ). If all  $\Delta Q$ 's are negative the algorithm halts.

After making a change the algorithm iterates. The algorithm halts when no  $\Delta v$  produces a  $\Delta Q > 0$ . In practice this means computing all these possibilities at each iteration, consuming a significant amount of processing time. A common compromise is to rank the jobs (most likely to improve  $Q$  the most at the top) and compute only the variable changes for one job at a time, take the best of those, and move to the next job, etc.

Keep in mind that a *move* could cause a constraint violation. Such a violation can come from any of the following:

- window constraints: each job $_i$  has a window of opportunity  $[w_{0i}, w_{1i}]$ , and we cannot schedule job $_i$  outside this window.
- duration constraints: each job $_i$  has a minimum and maximum duration  $[d_{\min i}, d_{\max i}]$ , and the scheduled duration ( $d_i$ ) must satisfy  $d_{\min i} \leq d_i \leq d_{\max i}$ .
- capacity constraints: each time unit can only have one job assigned to it (i.e.: resource capacity = 1).

Moves that violate the window and duration constraints are considered illegal and are not tolerated. The three algorithms that we describe below are distinguished by how they handle a potential capacity violation. For example if job $_i$  has an immediate neighbor on the right (e.g.: no time gap between the end of job $_i$  and the beginning of job $_j$ ) then growing job $_i$  one time unit to the right will cause a conflict (i.e.: job $_i$  *collides*



with job<sub>j</sub>). We have a few options at this point: consider the move illegal or try to move job<sub>j</sub> to make room for the move. The simplest algorithm is to freeze all the other jobs when we try to move a given job, that is, make no attempt to adjust neighbors to accommodate moves. Below, we consider three algorithms, beginning with the simplest.

**Simple Hill Climbing Algorithm:** For all jobs on the schedule consider all the possible legal moves (1 time unit). A move that violates any of the constraints (window, duration, capacity) is considered illegal. Take the legal move that increases Q the most, then iterate. The algorithm halts when there is no legal move with  $\Delta Q > 0$ . Thus, there are 6 moves to consider for each job<sub>i</sub>:

1. grow\_R(i)
2. grow\_L(i)
3. slide\_R(i)
4. slide\_L(i)
5. shrink\_R(i)
6. shrink\_L(i).

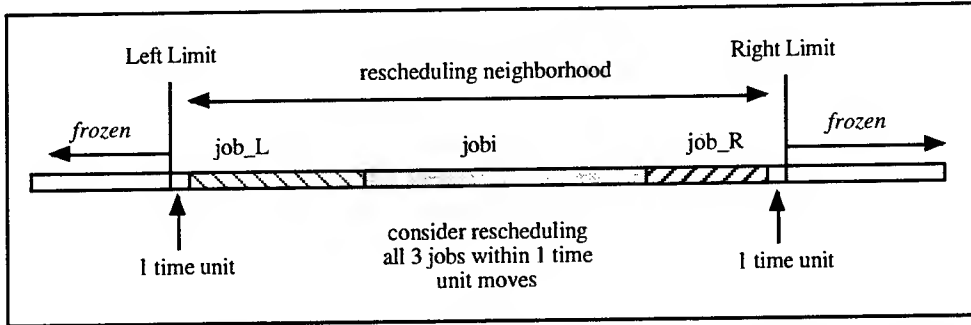
**Local Pressure Algorithm:** This algorithm is the same as the "simple algorithm" except that we consider making room for a move if a collision occurs. If job<sub>i</sub> collides with job<sub>j</sub> then we consider sliding and/or shrinking job<sub>j</sub> to accommodate the move. Thus, there are 14 possible moves, the first 6 are when there is no collision and the last 8 are when job<sub>i</sub> collides with job<sub>j</sub>:

1. grow\_R(i)
2. grow\_L(i)
3. slide\_R(i)
4. slide\_L(i)
5. shrink\_R(i)
6. shrink\_L(i)
7. grow\_R(i) & slide\_R(j)
8. grow\_R(i) & shrink\_L(j)
9. grow\_L(i) & slide\_L(j)
10. grow\_L(i) & shrink\_R(j)
11. slide\_R(i) & slide\_R(j)
12. slide\_R(i) & shrink\_L(j)
13. slide\_L(i) & slide\_L(j)
14. slide\_L(i) & shrink\_R(j).

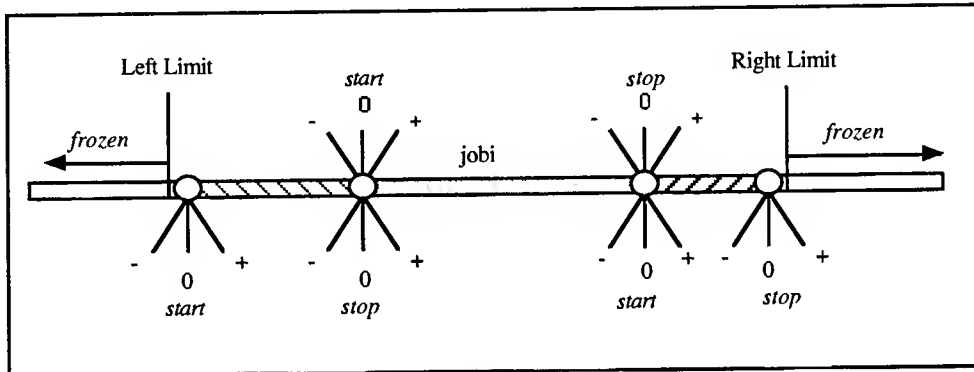
If it is not possible to move job<sub>j</sub> without causing another constraint violation then the move is not legal. We could say that the local pressure algorithm only goes "one deep" since it only considers moving an immediate neighbor. We could easily extend the algorithm to consider going "two deep" by recursively applying the same moves to the job that job<sub>j</sub> collides with, etc. For the local pressure algorithm we limit the *ripple* to the immediate neighbors for now.

**Neighborhood Search (Radius = 1):** For each job<sub>i</sub>, consider its immediate neighbors. An immediate neighbor is a job that has no time gap between it and the start or stop of job<sub>i</sub>. There can be either 2, 1 or 0 immediate neighbors. If there are 0 immediate neighbors we consider the 6 simple moves, as above.

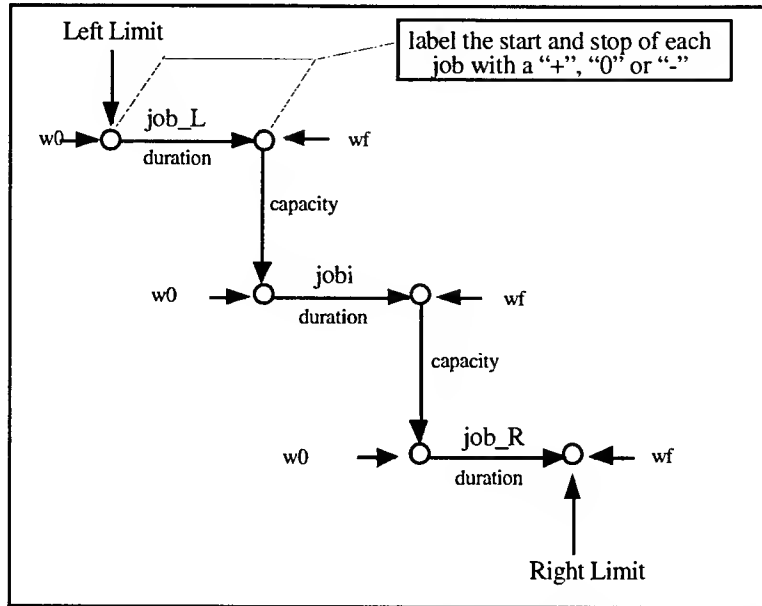
If there are 2 immediate neighbors, then there is one on the left (job\_L) and one on the right (job\_R). Define a "window" of time for rescheduling these three jobs: job\_L, job<sub>i</sub>, job\_R. We freeze all other jobs on the schedule outside this window. The left limit of the window is the maximum of {start of job



**Figure 27:** For the job<sub>i</sub> we consider a neighborhood of radius 1. In the sketch job<sub>i</sub> has immediate neighbors on both sides. The rescheduling neighborhood is defined to be 1 time unit to the left and right of the neighbors respectively. If there is no gap to the left of job<sub>L</sub> then the Left Limit is taken to be the start of job<sub>L</sub>, and if there is no gap to the right of job<sub>R</sub> the Right Limit is taken to be the stop of job<sub>R</sub>. All jobs that are scheduled outside this neighborhood are considered frozen (i.e.: cannot be adjusted in any way).



**Figure 28:** Within the rescheduling window we consider only moves of 1 time unit. That is, the start and stop of a job can only be: increment ("+" means "add one"); leave the same ("0"); or decrement ("- means "subtract one"). All such adjustments must satisfy the constraints (window, duration, capacity). We can then view the problem as a "labeling" problem: label the start and stop of each of the 3 jobs with a "+", "0" or "-" and find the labeling that produced the highest  $\Delta Q$ . Notice that since the stop of job<sub>L</sub> is initially equal to the start of job<sub>i</sub>, there are only 6 ways to label them without causing a capacity violation. For example, if the stop of job<sub>L</sub> is labeled "+" then the only way to label the start of job<sub>i</sub> is "+". The maximum number of possible labeling is  $= 3^2 \cdot 6^{k-1}$  where  $k = 2 \cdot \text{Radius} + 1$ , and  $\text{Radius} = 1$ . Thus, the maximum number of labelings when  $\text{Radius} = 1$  is 324. Many of these labelings will be illegal because of window and duration constraints, and possibly because the Left and Right Limits do not allow "-" and "+" respectively.

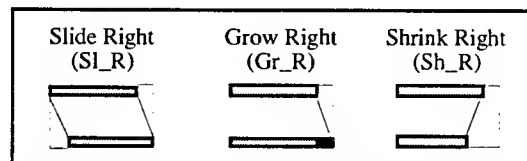


**Figure 29:** As we label the start and stop times, from left to right, the constraints propagate. For example the start of job\_L is constrained by the Left Limit and its left window limit (w0). When the start of job\_L is labeled then the duration constraints and right window constraint for job\_L constrain the labels for the stop of job\_L. Once the stop of job\_L is labeled then capacity constraints the label for the start of job\_i (and so does the left window limit for job\_i), etc.

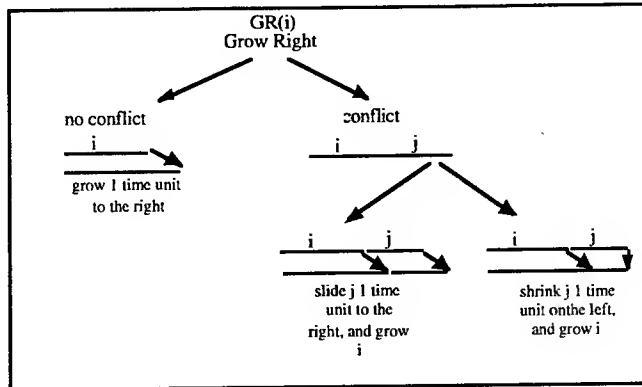
We do the same thing if job\_i has 0 on 1 immediate neighbors. For example, if job\_i has 0 immediate neighbors then we consider all the ways to move the start and stop by one time unit. This includes all the 6 moves described in the "simple" algorithm, as well as some different ones (e.g.: simultaneously shrinking on both the left and right), for a total of 9 possible moves.

Also notice that if job\_i and job\_j are immediate neighbors of each other, and have no other immediate neighbors, then the calculation that we do when we come to job\_i need not be repeated when we come to job\_j.

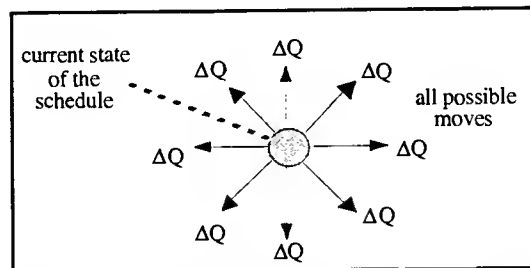
-----  
consider all the (small) adjustments to schedule parameters that might improve the schedule and choose the one that improves the schedule the most, and iterate.



**Figure 30:** Some of the "moves" that might be applied to a job that is on the schedule. Notice that the moves must obey the widow and duration constraints. The Slide and Grow moves might conflict with a neighboring job on the schedule, in which case the decision must be made to either not allow the move or accommodate it by modifying the neighbor (e.g.: slide, shrink or bump the neighbor).



**Figure 31:** Hill Climbing is complicated by the fact that an incremental change in one parameter may cause a ripple effect as it affects other jobs on the schedule. For example, growing a job to the right might conflict with a neighboring job, as shown in the figure (job *i* bumps into job *j*). Our algorithms limit the amount of ripple allowed: if we cannot accommodate the move by with a simple adjustment of the neighbor (slide or shrink the neighbor, but do not bump the neighbor off the schedule) we dis-allow the move.



**Figure 32:** Hill Climbing considers all possible incremental changes in the state of the schedule and then picks the one that increase  $Q$  by the most (largest  $\Delta Q$ ). If all  $\Delta Q$ 's are negative the algorithm halts.

We have explored a variety of hill climbing strategies (e.g.: bump vs. no-bump). The simplest version is to grow the jobs on the schedule (highest priority first) within the constraints (window and duration limits) and without moving any neighbors. This is the same as the "optimization" phase of the PD algorithm.

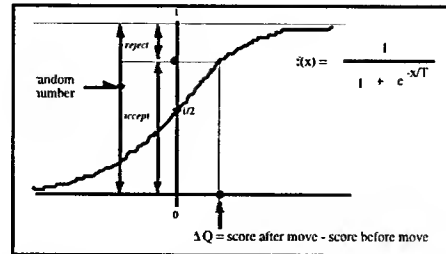
We have come to view the Hill Climbing algorithms as in either of two classes: simple enough that they are thought of as variations on the third phase of the PD algorithm; or as sufficiently complicated that they then form the backbone of both the Simulated Annealing and Tabu Search approaches. That is, the complications introduced by a sophisticated hill climber<sup>18</sup> will increase the accuracy of the schedules but at the expense of run time and complexity. We have not found this accuracy-speed-simplicity trade off to be worth it: the relatively small increase in accuracy comes at a high price. In fact, the Simulated Annealing and Tabu Search methods fall into the same category. Although they can provide significant improvements in accuracy, they also introduce significant increases in run time (several orders of magnitude longer than the PD) and complexity.

**Simulated Annealing:** This method begins the same way as the Hill Climber but adds a random component that makes random decisions in the beginning, and gradually less random decisions as it progresses, and finally converges to a deterministic Hill Climbing method.

Consider the moves that were possible for the Hill Climber, and add moves that might have been ignored because they were certain to produce a negative  $\Delta Q$ . At any state of the schedule we can randomly pick one

<sup>18</sup> For example, we might allow swapping position with the neighbor, and other such "local" moves.

of the moves and consider the  $\Delta Q$  associated with it. The following rule is applied<sup>19</sup>: map the value of  $\Delta Q$  through the sigmoidal function shown in the diagram below; this produces the value  $f(\Delta Q)$  which is some number between 0 and 1. Then we compare that number to a randomly generated number between 0 and 1 and use the rule: if the random number is greater than  $f(\Delta Q)$  then do not make the move, and iterate; if the random number is less than or equal to the  $f(\Delta Q)$  then make the move and iterate.



**Figure 33:** The "accept/reject" method used in Simulated Annealing depends on the shape of the sigmoid, which in turn is controlled by the value of the "temperature"  $T$ .

Notice that the steeper the curve the more likely we are to accept moves with positive  $\Delta Q$  and to reject moves with negative  $\Delta Q$  (i.e.: similar to the Hill Climber). Conversely, the flatter the curve the more likely we are to be randomly accepting and rejecting moves (independent of the value of  $\Delta Q$ ). The parameter " $T$ " is called the "temperature" and when the temperature is high the moves are random and when the temperature is low the algorithm behaving like the Hill Climber. Thus, SA requires a "temperature schedule" that usually begins with a high value of  $T$  and successively lowers it until it is small enough that Hill Climbing takes over and then halts when no improvements can be made to the schedule. Along the way, many schedules are "observed" so the method implies that we keep track of the best schedule that we have seen through the whole process.

The advantage to the SA approach is that it avoids getting stuck at local optima and can produce some of the best results for a wide range of NP-Complete problems<sup>20</sup>. The drawback is that it introduces randomness into the schedule making process so it may not be repeatable and it difficult to understand "how" the result was arrived at (in layman's terms). The major drawback, however, is that the SA results tend to come at the expense of significant increases in run time (we have ruled this method out for this reason).

It is worth mentioning at this point that an "agent-based" approach could use the same "moves" while viewing each job as an "agent" that "looks" at its environment (competing jobs, neighbors) and makes a "move". In this view the "jobs" are interacting like a community of agents, observing their environment, communicating, negotiating, etc. We have only explored limited versions of such an approach, and in fact, ended up implementing simulated annealing in its place.

**Tabu Search:** We did not simulate a Tabu Search approach, and although we plan to test such a method in the future, we extrapolated from our experience with Hill Climbing, Simulated Annealing, and Genetic Algorithms to come to the conclusion that it will introduce unacceptable increases in run time and unacceptable complexity. Tabu Search does, however, show promise in that it minimizes one of the major weaknesses of the SA approach (it removes the randomness of the decisions), while exploiting its strength (visit local optima, but do not get stuck there). Our fear is that it will lead to a *backtracking* search, however controlled and intelligent it might be. It is our experience that backtracking must be introduced in

<sup>19</sup> In some SA applications the following rule is applied: if  $\Delta Q > 0$  then make the move (no randomness) and iterate.

<sup>20</sup> For many problems where the optimal solution is unknown, the "best known" results are almost always gotten with a SA approach, and they tend to be used for comparisons.

extremely limited ways or the run times quickly become unacceptable. This intuition, combined with the simplicity issue, points us away from the Tabu Search method. But in all fairness this is an open question.

# Development of a New Numerical Boundary Condition for Perfect Conductors

Jeffrey L. Young  
Associate Professor  
Department of Electrical Engineering

University of Idaho  
Moscow, OH

Final Report for:  
Summer Research Extension Program  
Wright Laboratory

Sponsored by:  
Air Force Office of Scientific Research  
Bolling AFB, Washington, DC

and

Wright Laboratory

December 1996

---

## DEVELOPMENT OF A NEW NUMERICAL BOUNDARY CONDITION FOR PERFECT CONDUCTORS

Jeffrey L. Young  
Associate Professor  
Department of Electrical Engineering  
University of Idaho

### Abstract

A new numerical boundary condition is derived that extends Maxwell's equations to the surface of a perfect conductor. This condition explicitly shows the interrelationship between surface charge, surface current and tangential electric field. Particularly, the boundary equations are similar to Euler's linearized inviscid acoustic equations, where current density is akin to fluid velocity and charge density is akin to mass density; the normal derivative of the tangential electric field is the source term in the equivalent momentum equation.

The two sets of equations, Maxwell's and Euler's, are discretized using the finite-volume, MUSCL procedure along with the two-stage Runge-Kutta integrator. Specifically, the spatial discretization employs the flux-splitting procedure along windward-biased differencing. Such an approach captures the solution within its domain of influence, as required from a solution of a set of hyperbolic equations. Formally, the resulting scheme is third-order accurate in space and second-order accurate in time.

Numerical results are provided to validate the proposed methodology by considering plane wave scattering from a perfectly conducting sphere. Such an example provides all the necessary complexity to exercise the algorithm. Numerical data are provided that show the progression of charge density on the sphere and the resulting radar cross-section (RCS) produced by the sphere. Data comparisons are made with another method that employs a first-order type boundary condition. For the RCS data, the theoretical Mie series is used for benchmarking purposes.

Although the data shows that the scheme works, there are problems with the scheme in terms of late-time instability. The causes of this instability are currently under investigation and are the subject of a future study.



# DEVELOPMENT OF A NEW NUMERICAL BOUNDARY CONDITION FOR PERFECT CONDUCTORS

Jeffrey L. Young

## 1 Introduction

Over the past decades, several time-domain methods for Maxwell's equations have appeared in the literature. The most popular is the finite-difference, time-domain (FDTD) method of Yee [1],[2]. The power of this method rests on its algorithmic simplicity and its ability to discretize Maxwell's equations to second-order. Although the method has been used in many diverse electromagnetic applications, its primary shortcoming is its inability to form a grid that naturally conforms to the body under investigation.

To circumvent this problem, Madsen [3] developed a modified body-fitting FDTD method that preserves the divergence properties of the electromagnetic vectors. Despite its use of the dual grid and its problem with late time instabilities, it has been shown to produce accurate results in several applications. Edge elements have recently replaced the traditional nodal-based elements as the elements of choice in the weighted-residual paradigm (see Mahadevan and Mittra [4] for a review of the edge-element methodology). These too admit no spurious numerical charge since the element is customized to satisfy the divergence equations of Maxwell. As with the method of Madsen and Yee, the unknowns in the edge-element algorithm are not collocated. From a grid generation and a coding perspective, this uncollocated strategy adds significant complexity to the problem.

More recently, the finite-volume approach, which was developed for the fluid dynamics community, has been considered for electromagnetic-type problems. For this situation, Maxwell's equations are couched in strong conservative form and the unknowns are collocated at the cell centers. Flux evaluations at the cell walls are determined from an interpolation of the dependent variables at cell centers. For certain types of interpolative schemes and flux-splitting procedures, the algorithm mimics many of the features of the hyperbolic system of equations. For example, when the fluxes are split according to the eigenvalues of the flux Jacobian matrix and windward-based differencing is used in the discretization, the algorithm naturally captures the right- and left- running waves within their domain of influence.

The main contributors to the finite-volume strategy, as applied to Maxwell equations, are Shankar of Rockwell [5, 6, 7] and Shang of the Wright Patterson Air Force Base [8, 9, 10]. The method of Shankar

is founded on the upwind Lax-Wendroff method. For example, when applied to the model scalar wave equation, the solution  $u$  at time  $n + 1$  and node  $j$  is derived from:

$$u_j^{n+1} = u_j^n + .5\nu(3u_j^n - 4u_{j-1}^n + u_{j-2}^n) + .5\nu^2(u_{j+1}^n - 2u_j^n + u_{j-1}^n),$$

where  $\nu = c\delta t/\delta x$ . When applied to higher dimensions, Shankar decomposes the dependent variable in terms of two states, from which he derives the flux at the cell interface. Shang also decomposes the dependent variable in terms of two states, but the flux evaluations are obtained from a flux-splitting procedure and the modified MUSCL algorithm [11] (sometimes referred to as the  $\kappa$ -scheme). By suitably selecting the appropriate values for  $\kappa$  and the limiter  $\phi$ , the algorithm can take on one of five different spatial discretization schemes: first-order windward, second-order windward, second-order Fromm, third-order windward biased or second-order central differenced.

Unfortunately, there exist inherent difficulties in maintaining high-order accuracy when these methods are applied near and on perfect conductors. Two fundamental problems prevent one from establishing anything more than a first-order approximation to the boundary conditions. First, all field components (magnetic and electric) share the same point in space. If that point rests on the perfect conductor, then only the values of three of the six components are known *a priori* on the conductor; these values are tangential electric field and normal magnetic field. The other three, normal electric field and tangential magnetic field are not known since the former is proportional to the induced surface charge and the latter is proportional to the induced surface current density. To deduce the values for the charge and the current, one possible approach is to extrapolate the normal electric field and tangential magnetic field from interior values. Second, when second-order (or greater) windward differencing is used, at least two cell values on either side of the flux wall are required to reconstruct the flux on the wall. For cell walls one cell away from the conductor, one of these cell values lies in the interior of the conductor, which is technically outside the computational domain. This being the case, a first-order approximation is required.

In this report, we present a new strategy that is locally and globally third-order accurate in space. This is achieved by reconsidering and manipulating Maxwell's equations in the vicinity of the conductor. Particularly, the result of this analysis is a set of two time-domain equations that explicitly show the interrelationship between surface charge density, surface current density and tangential electric field. As might be expected, the interrelationship is quite similar to Euler's linearized, inviscid acoustic equations, with the tangential field acting like a source of "current momentum."

The numerical procedure rests upon the simultaneous discretization of both Maxwell's and Euler's equa-

tions. In this report the discretization will be accomplished via the MUSCL procedure in generalized curvilinear coordinates. To advance the equations in time, the second-order Runge-Kutta (RK) scheme is employed. For sake of completeness, the scheme is numerically validated by considering the scattering of a TEM plane wave from a perfectly conducting sphere. To this end, plots of the sphere's radar cross-section (RCS) and the surface charge are provided.

## 2 Governing Equations

Consider Maxwell's curl equations for linear, isotropic, homogeneous media:

$$\mu \frac{\partial \mathbf{H}}{\partial t} = -\nabla \times \mathbf{E} \quad (1)$$

and

$$\epsilon \frac{\partial \mathbf{E}}{\partial t} = \nabla \times \mathbf{H}, \quad (2)$$

where  $\mathbf{E}$  and  $\mathbf{H}$  are the electric and magnetic fields, respectively;  $\epsilon$  and  $\mu$  are the permittivity and permeability, respectively. Let  $\mathbf{n}$  be a constant unit vector normal to and pointing out from the perfect conductor. Then, by taking the scalar product of  $\mathbf{n}$  with Ampere's law, we obtain

$$\epsilon \mathbf{n} \cdot \frac{\partial \mathbf{E}}{\partial t} = \nabla \cdot (\mathbf{H} \times \mathbf{n}) + \mathbf{H} \cdot (\nabla \times \mathbf{n}), \quad (3)$$

where it is understood that the above equation is being considered at an infinitesimal distance from the perfect conductor. Since the curl of a constant vector is zero,  $\mathbf{n} \times \mathbf{H}$  is the surface current  $\mathbf{J}_s$ , and  $\epsilon \mathbf{n} \cdot \mathbf{E}$  is the surface charge  $\rho_s$ , Eqn. (3) is equivalent to

$$\frac{\partial \rho_s}{\partial t} + \nabla \cdot \mathbf{J}_s = 0. \quad (4)$$

As expected, the resultant equation is simply a statement that charge must be conserved on a perfect conductor. By way of comparison, we observe the analogous relationship between the acoustic quantities of fluid velocity and mass density with the electromagnetic quantities of current density and charge density, respectively.

The previous derivation is duplicated by forming the vector product of  $\mathbf{n}$  with Faraday's law:

$$\mu \mathbf{n} \times \frac{\partial \mathbf{H}}{\partial t} = -\nabla(\mathbf{n} \cdot \mathbf{E}) + (\mathbf{n} \cdot \nabla)\mathbf{E} + (\mathbf{E} \cdot \nabla)\mathbf{n} + \mathbf{E} \times (\nabla \times \mathbf{n}) \quad (5)$$

Of course, the last two terms on the left-hand side are zero due to the requirement that  $\mathbf{n}$  be a constant. More importantly, the remaining term on the right-hand side is nothing more than the rate of change of

surface current and the first term on the left-hand side is proportional to the gradient of charge density.

Thus, we are left with

$$\frac{\partial \mathbf{J}_s}{\partial t} + v^2 \nabla \rho_s = (\mathbf{n} \cdot \nabla) \mathbf{E} / \mu, \quad (6)$$

where  $v = 1/\sqrt{\mu\epsilon}$  is the phase velocity of the homogeneous medium. The vector senses of  $\mathbf{J}_s$  and  $\nabla \rho_s$  are tangentially directed to the conductor's surface, which implies that the source term must also be directed similarly. Hence,

$$\frac{\partial \mathbf{J}_s}{\partial t} + v^2 \nabla \rho_s = \frac{1}{\mu} \frac{\partial \mathbf{E}_t}{\partial n}, \quad (7)$$

where  $\mathbf{E}_t$  is the tangential component of  $\mathbf{E}$ . Here we see the analogous momentum conservation equation whose source term is the normal derivative of tangential  $\mathbf{E}$ .

To summarize, the governing equations for the domain variables  $\mathbf{E}$  and  $\mathbf{H}$  are stated in Maxwell's equations, (1) and (2). The governing equations for the boundary variables  $\mathbf{J}_s$  and  $\rho_s$  are stated in Euler's equations, (4) and (7). Finally, the system is made complete via the boundary conditions:

$$\mathbf{n} \times \mathbf{E} = 0 \quad (8)$$

$$\mathbf{n} \cdot \mathbf{E} = \rho_s / \epsilon \quad (9)$$

$$\mathbf{n} \times \mathbf{H} = \mathbf{J}_s, \quad (10)$$

and

$$\mathbf{n} \cdot \mathbf{H} = 0. \quad (11)$$

These systems of equations (Maxwell, Euler and boundary conditions) allow one to compute both boundary variables and domain variables simultaneously. Consequently, a numerical procedure can be developed to utilize that fact, thus circumventing the problem of not knowing the surface current and charge densities on the conductor *a priori*. Two other comments are in order. First, it is well known that tangential current leads to electromagnetic radiation. By treating  $\mathbf{J}_s$  as an dependent variable to be computed, we see that the derived Euler's equations have a primary importance of computing  $\mathbf{J}_s$ , and that Maxwell's equations are needed only for the deduction of the source field in the analogous momentum equation. Second, the source term in the momentum equation is easily discretized by noting that the node that resides on the conductor is zero, since tangential  $\mathbf{E}$  must vanish on the conductor.

## 2.1 Conservative Form

We begin this treatment by couching Maxwell's equations in the following conservative form:

$$\frac{\partial U}{\partial t} + \nabla \cdot (\mathbf{F}(U)) = 0, \quad (12)$$

where  $U$  is the solution vector and  $\mathbf{F}$  is the point-flux tensor, which is dependent upon  $U$ . In expanded form,

$$\frac{\partial U}{\partial t} + \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z} = 0, \quad (13)$$

where  $F_x, F_y, F_z$  are the flux components of  $F$ . Since the the fluxes are homogeneous functions of degree one, we may write  $F_x = AU$ ,  $F_y = BU$  and  $F_z = CU$ . Here  $A, B, C$  are the Jacobian matrices associated with  $F_x, F_y, F_z$ , respectively. Thus, Equation (13) is equivalent to

$$\frac{\partial U}{\partial t} + \frac{\partial(AU)}{\partial x} + \frac{\partial(BU)}{\partial y} + \frac{\partial(CU)}{\partial z} = 0. \quad (14)$$

For homogeneous, isotropic media, it is a simple exercise to show that the flux Jacobian matrices associated with Maxwell's equations are given by

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -\frac{1}{\epsilon} \\ 0 & 0 & 0 & 0 & \frac{1}{\epsilon} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\mu} & 0 & 0 & 0 \\ 0 & -\frac{1}{\mu} & 0 & 0 & 0 & 0 \end{bmatrix}, \quad (15)$$

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \frac{1}{\epsilon} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\frac{1}{\epsilon} & 0 & 0 \\ 0 & 0 & -\frac{1}{\mu} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{\mu} & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad (16)$$

and

$$C = \begin{bmatrix} 0 & 0 & 0 & 0 & -\frac{1}{\epsilon} & 0 \\ 0 & 0 & 0 & \frac{1}{\epsilon} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -\frac{1}{\mu} & 0 & 0 & 0 & 0 \\ \frac{1}{\mu} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad (17)$$

where

$$U = [B_x, B_y, B_z, D_x, D_y, D_z]^t. \quad (18)$$

Hence,

$$F_x = [0, -D_z/\epsilon, D_y/\epsilon, 0, B_z/\mu, -B_y/\mu]^t \quad (19)$$

$$F_y = [D_z/\epsilon, 0, -D_x/\epsilon, -B_z/\mu, 0, B_x/\mu]^t \quad (20)$$

and

$$F_z = [-D_y/\epsilon, D_x/\epsilon, 0, B_y/\mu, -B_x/\mu, 0]^t. \quad (21)$$

Note: From the constitutive relationships,  $\mathbf{B} = \mu\mathbf{H}$  and  $\mathbf{D} = \epsilon\mathbf{E}$ .

Except for the inclusion of the source term (i.e., the normal derivative of tangential  $\mathbf{E}$ ), the same representations given by Eqn. (13) or (14) may be used for Euler's equations. Particularly, for the flux Jacobian matrices,

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ v^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad (22)$$

$$B = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ v^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad (23)$$

and

$$C = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ v^2 & 0 & 0 & 0 \end{bmatrix}; \quad (24)$$

for the dependent variables,

$$U = [\rho_s, J_x, J_y, J_z]^t. \quad (25)$$

From these definitions it follows that

$$F_x = [J_x, v^2 \rho_s, 0, 0]^t \quad (26)$$

$$F_y = [J_y, 0, v^2 \rho_s, 0]^t \quad (27)$$

and

$$F_z = [J_z, 0, 0, v^2 \rho_s]^t. \quad (28)$$

## 2.2 Flux Splitting

The flux Jacobian matrix  $A$  (as well as  $B$  and  $C$ ) associated with both systems is diagonalizable since its eigenvectors are linearly independent. That is, an alternative representation for  $A$  is

$$A = S_x \Lambda_x S_x^{-1} \quad (29)$$

where  $S_x$  is the modal matrix of  $A$  and  $\Lambda_x$  is its spectral matrix, which is a diagonal matrix. A detailed eigenvector analysis reveals for Maxwell's equations that

$$\Lambda_x = \text{Diag} \{-\lambda, -\lambda, \lambda, \lambda, 0, 0\}, \quad (30)$$

where  $\lambda = 1/\sqrt{\mu\epsilon}$ ; also

$$S_x = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ \eta & 0 & -\eta & 0 & 0 & 0 \\ 0 & -\eta & 0 & \eta & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}, \quad (31)$$

with  $\eta = \sqrt{\mu/\epsilon}$ .

The flux splitting procedure is based upon the decomposition of the spectral matrix in terms of the sign of its members:

$$\Lambda_x = \Lambda_x^+ + \Lambda_x^- \quad (32)$$

where

$$\Lambda_x^+ = \text{Diag} \{0, 0, \lambda, \lambda, 0, 0\} \quad (33)$$

and

$$\Lambda_x^- = \text{Diag} \{-\lambda, -\lambda, 0, 0, 0, 0\}. \quad (34)$$

It follows from Eqn. (29) that

$$A = S_x(\Lambda_x^+ + \Lambda_x^-)S_x^{-1} = A^+ + A^-. \quad (35)$$

Here  $A^+ = S_x\Lambda_x^+S_x^{-1}$  and  $A^- = S_x\Lambda_x^-S_x^{-1}$ . Since  $F_x = AU$ , then  $F_x = (A^+ + A^-)U = F_x^+ + F_x^-$ , where  $F_x^+ = A^+U$  and  $F_x^- = A^-U$ . Carrying out the required matrix operations, we discover that

$$F_x^+ = \frac{1}{2} \left[ 0, \left( B_y \lambda - \frac{D_z}{\epsilon} \right), \left( B_y \lambda + \frac{D_y}{\epsilon} \right), 0, \left( \frac{B_z}{\mu} + D_y \lambda \right), \left( \frac{-B_y}{\mu} + D_z \lambda \right) \right]^t \quad (36)$$

and

$$F_x^- = \frac{1}{2} \left[ 0, -\left( B_y \lambda + \frac{D_z}{\epsilon} \right), \left( -B_y \lambda + \frac{D_y}{\epsilon} \right), 0, \left( \frac{B_z}{\mu} - D_y \lambda \right), -\left( \frac{B_y}{\mu} + D_z \lambda \right) \right]^t \quad (37)$$

By way of a similar analysis, the split fluxes for Euler's equations are found to be

$$F_x^+ = \frac{1}{2} [v\rho_s + J_x, v^2\rho_s + vJ_x, 0, 0]^t \quad (38)$$

and

$$F_x^- = \frac{1}{2} [-v\rho_s + J_x, v^2\rho_s - vJ_x, 0, 0]^t. \quad (39)$$

Regardless of the governing equations to be considered, the duplication of the flux splitting procedure for the other fluxes yields the expanded form for the conservative equation:

$$\frac{\partial U}{\partial t} + \frac{\partial F_x^+}{\partial x} + \frac{\partial F_x^-}{\partial x} + \frac{\partial F_y^+}{\partial y} + \frac{\partial F_y^-}{\partial y} + \frac{\partial F_z^+}{\partial z} + \frac{\partial F_z^-}{\partial z} = 0. \quad (40)$$

The previous result is the equation on which characteristic theory is based. A simple discretization scheme is given next to demonstrate the basic concepts of the numerical theory.

## 2.3 A Simple Difference Scheme

The advantage of writing Maxwell's or Euler's equations in characteristic form, as manifested in Eqn. (40), is found in the differencing scheme of the spatial derivatives. Since  $F_x^+$  corresponds to eigenvalues that are positive, which leads to positive going waves in the x-direction in a one-dimensional analysis, it makes sense to apply a windward differencing scheme to  $F_x^+$ . For second-order accuracy, we may use

$$\left. \frac{\partial F_x^+}{\partial x} \right|_i = \frac{3F_i^+ - 4F_{i-1}^+ + F_{i-2}^+}{2\delta_x} \quad (41)$$

Hence, it is readily seen that the wave is discretized solely in terms of its values within its domain of influence.

Similarly,  $F^-$  is associated with waves that are traveling in the negative x-direction. Hence, we let

$$\left. \frac{\partial F_x^-}{\partial x} \right|_i = \frac{3F_i^- - 4F_{i+1}^- + F_{i+2}^-}{2\delta_x}. \quad (42)$$

Note: The use of windward differencing will introduce a certain amount of numerical dissipation and dispersion into the solution.

For certain classes of temporal integrators, such as the Runge-Kutta integrators, the previous scheme is found to be conditionally stable. Formally, the RK methods may be expressed in terms of a truncated Taylor series for the matrix  $e^{-A\delta_t}$ , where  $A$  is the spatial discretization matrix:

$$f^{n+1} = \sum_{m=1}^M \frac{(-\delta_t A)^m}{m!} \cdot f^n. \quad (43)$$

Obviously, the implementation of this series is straightforward and any order of accuracy can be achieved, if one is willing to expend the computational resources to achieve that accuracy. However, to minimize the number of evaluations of the right-hand side of Eqn. (43) and to limit the storage required for each evaluation, the second-order and fourth-order schemes are used. For second-order accuracy,

$$\begin{aligned} k_0 &= \delta_t R(f_0) \\ k_1 &= \delta_t R(f_1) \\ f^{n+1} &= f^n + k_1. \end{aligned} \quad (44)$$

Here  $R$  is the residual;  $f_0 = f(x, t_0)$  and  $f_1 = f_0 + k_0/2$ . For fourth-order accuracy,

$$\begin{aligned} k_0 &= \delta_t R(f_0) \\ k_1 &= \delta_t R(f_1) \\ k_2 &= \delta_t R(f_2) \end{aligned}$$



$$\begin{aligned}
k_3 &= \delta_t R(f_3) \\
f^{n+1} &= f^n + (k_1 + 2k_2 + 2k_3 + k_4)/6.
\end{aligned} \tag{45}$$

Here  $R$  is the residual;  $f_0 = f(x, t_0)$ ,  $f_1 = f_0 + k_0/2$ ,  $f_2 = f_1 + k_1/2$  and  $f_3 = f_2 + k_2$ . A detailed Fourier analysis shows that the second-order, fully upwind, two-stage Runge-Kutta scheme is conditionally stable with a CFL=.5; for the four-stage scheme, CFL=.695.

To apply the flux-splitting procedure to a grid that is non-Cartesian, additional analytical work is required. This subject is studied next.

## 2.4 Strong Conservative Form

To cast the conservative equation into strong conservative form, we map the Cartesian independent variables  $x, y, z$  into a generalized coordinate system  $\xi, \eta, \zeta$ ; the dependent variables are still expressed in terms of the Cartesian frame. Hence, consider the following coordinate transformation:  $\xi = \xi(x, y, z)$ ,  $\eta = \eta(x, y, z)$  and  $\zeta = \zeta(x, y, z)$ ; or similarly, for one-to-one transformations,  $x = x(\xi, \eta, \zeta)$ ,  $y = y(\xi, \eta, \zeta)$  and  $z = z(\xi, \eta, \zeta)$ . Associated with this transformation is a set of metrics (e.g.,  $x_\eta$ ,  $\eta_z$ ) that convey the geometry of the transformation.

After a sequence of mathematical operations that involve the metrics of the transformation, Eqn. (13) becomes

$$\frac{\partial \hat{U}}{\partial t} + \frac{\partial \hat{F}}{\partial \xi} + \frac{\partial \hat{G}}{\partial \eta} + \frac{\partial \hat{H}}{\partial \zeta} = 0, \tag{46}$$

where  $\hat{U}$  is the unknown domain vector and  $\hat{F}$ ,  $\hat{G}$ , and  $\hat{H}$  are the domain fluxes in the  $\xi$ ,  $\eta$ , and  $\zeta$  directions, respectively. With  $V$  denoting the Jacobian, the fluxes,  $\hat{U}$  and  $\hat{J}$  take on the following meaning:

$$\hat{U} = UV, \tag{47}$$

$$\hat{F} = (\xi_x F_x + \xi_y F_y + \xi_z F_z)V, \tag{48}$$

$$\hat{G} = (\eta_x F_x + \eta_y F_y + \eta_z F_z)V, \tag{49}$$

and

$$\hat{H} = (\zeta_x F_x + \zeta_y F_y + \zeta_z F_z)V. \tag{50}$$

A similar representation holds for Euler's equations, except that one term in the strong conservative equation may be omitted, since current flow is restricted to a locally two-dimensional surface.

### 3 Finite-Volume Procedure

The discretization of the conservative equation is accomplished via a cell-centered, finite-volume approach in conjunction with the Runge-Kutta two-stage integrator. By following the lead of Steger and Warming [12], we reconsider the split fluxes (e.g.,  $F^+$  and  $F^-$ ). These fluxes are associated with the right and left running waves at the cell interface; the states  $U^L$  and  $U^R$  are constructed from known values of  $U$  in adjacent cells. For example, the  $\kappa$ -scheme [11] requires that at interface  $i + 1/2$ ,

$$U_{i+1/2}^L = U_i + \frac{\phi}{4}[(1 - \kappa)\nabla + (1 + \kappa)\Delta]U_i,$$

and

$$U_{i+1/2}^R = U_{i+1} - \frac{\phi}{4}[(1 + \kappa)\nabla + (1 - \kappa)\Delta]U_{i+1},$$

where  $\phi$  is a limiter ( $\phi = 0, 1$ ),  $\kappa$  is an accuracy parameter,  $\nabla U_i = U_i - U_{i-1}$  and  $\Delta U_i = U_{i+1} - U_i$ . When  $\phi$  equals unity and  $\kappa$  takes on a value of  $-1, 1/3$  or  $+1$ , the scheme is deemed second-order windward, third-order windward-biased or second-order central differenced, respectively; the second-order Fromm scheme is recovered when  $\phi = 1$  and  $\kappa = 0$ . Thus, one can see that the advantage of the  $\kappa$ -scheme is found in its ability to capture the specific physics of the problem at hand. Although third-order accuracy is desired in most situations, a second-order windward scheme has the advantage of predicting the slope of the field at a dielectric interface from field values totally resident in the dielectric. In contrast, a central difference approximation will lead to boundary errors since the prediction of the dependent variable requires knowledge of the dependent variable on both sides of the dielectric interface; for large disparities between the interfacial permittivities, the discontinuities in either normal  $\mathbf{E}$  or tangential  $\mathbf{D}$  will not be captured. To predict these discontinuities correctly, two options exist: 1) use a fully windward scheme, which will require a smaller time step, as noted from the stability analysis [13] or 2) set  $\phi$  to a value of zero. For this latter case, a certain amount of numerical dissipation will be introduced into the solution due to the resulting first-order approximation.

Once the left and right states of  $U$  are estimated at the  $i + 1/2$  interface, the flux crossing that surface is simply,

$$F_{x,i+1/2} = F_x^+(U_{i+1/2}^L) + F_x^-(U_{i-1/2}^R).$$

Similarly, at interfaces  $j$  and  $k$ :

$$F_{y,j+1/2} = F_y^+(U_{j+1/2}^L) + F_y^-(U_{j-1/2}^R)$$

and

$$F_{z,k+1/2} = F_z^+(U_{k+1/2}^L) + F_z^-(U_{k-1/2}^R).$$

For generalized coordinates, the flux, say  $\hat{F}$ , is split in the direction of  $\xi$ . This is accomplished by a local transformation matrix  $T$  and by defining a new flux  $\bar{F}$ :  $\bar{F} = T\hat{F}$ . Since  $T$  can be chosen such that  $\bar{F}$  and  $\hat{F}$  have the same functional form, the eigenvalue/eigenvector information is directly obtained from the Cartesian formulation. Thus, if  $\bar{F} = \bar{F}^+ + \bar{F}^-$  then  $\hat{F} = \hat{F}^+ + \hat{F}^-$ , where  $\hat{F}^+ = T^{-1}\bar{F}^+$  and  $\hat{F}^- = T^{-1}\bar{F}^-$ . This same procedure is repeated in the directions  $\eta$  and  $\zeta$ , for  $\hat{G}$  and  $\hat{H}$ , respectively.

To maintain stability, the two-stage Runge-Kutta integrator is invoked, as defined earlier. Although other temporal schemes may be used, the RK integrator is simple to code and to vectorize for supercomputing platforms.

## 4 Numerical Results

To demonstrate some of the previous concepts, the problem of a plane wave scattering from a perfectly conducting sphere is considered. The computational grid is constructed by employing the Cartesian to Spherical coordinate mapping functions. This mapping results in a structured grid whose indices  $I, J, K$  denote the number of grid lines in the  $R, \theta, \phi$  directions, respectively. The sphere is illuminated from above by a TEM gaussian plane wave of the form  $e^{-w^2(t-z)^2}$ , where it has been assumed that the surrounding medium has unity permittivity and permeability, thus rendering the phase velocity to unity. To achieve a usable spectral content of the incident wave up to 12 r/s,  $w$  is set to 3.96. Finally, a CFL number of unity has experimentally been proven to yield stable results<sup>1</sup>.

Figures 1, 2 and 3 show plots of the charge density on the sphere as a function of elevation angle after 150, 300 and 450 time steps, respectively; for the present situation  $\Delta_R = 0.05$  m,  $(I, J, K) = (30, 30, 54)$  and  $\phi = 0$ . The data associated with one curve is obtained via a first-order boundary condition (by first-order, we mean that the unknowns that exist on the spherical surface are extrapolated from the interior via a first-order shift from the nearest cell); the other curve is associated with the exact boundary condition described in this report. These plots serve to validate the theory of the exact boundary condition. That is, with these plots confidence is gained from numerical experimentation that the charge density indeed satisfies the linearized inviscid Euler equations.

Using this same coarse grid and 1000 time steps, the RCS of the sphere is computed. Consider Figures

<sup>1</sup>Technically, the CFL number should be set to a value no greater than 0.87 [13]. However, due to the way the time step is calculated from the grid geometry, a CFL of 1.00 still results in stable data.

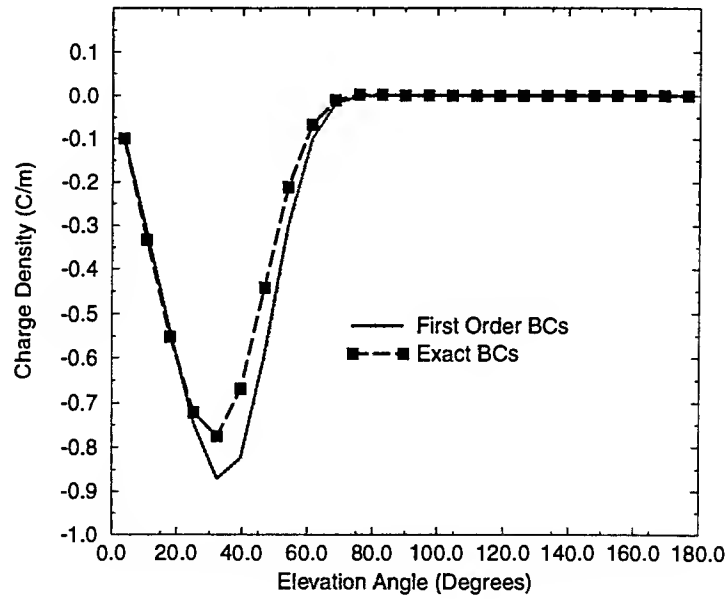


Figure 1: Charge density as a function of elevation angle after 150 time steps;  $\phi = 0$ .

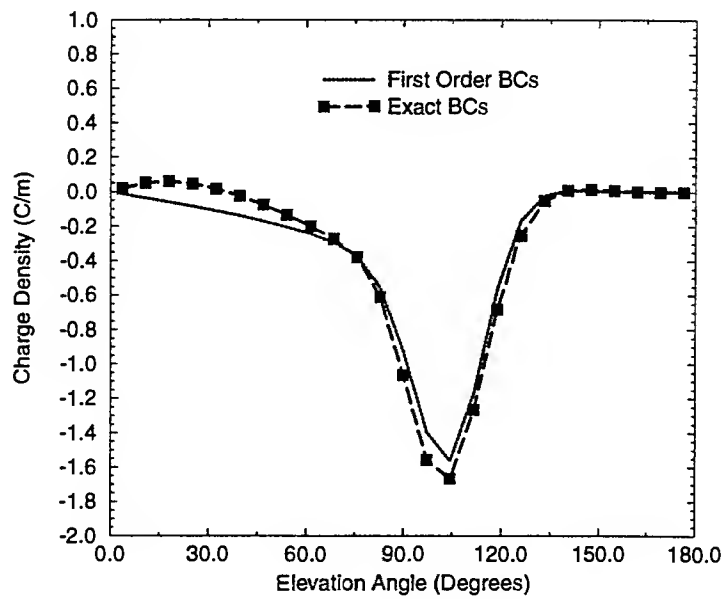


Figure 2: Charge density as a function of elevation angle after 300 time steps;  $\phi = 0$ .

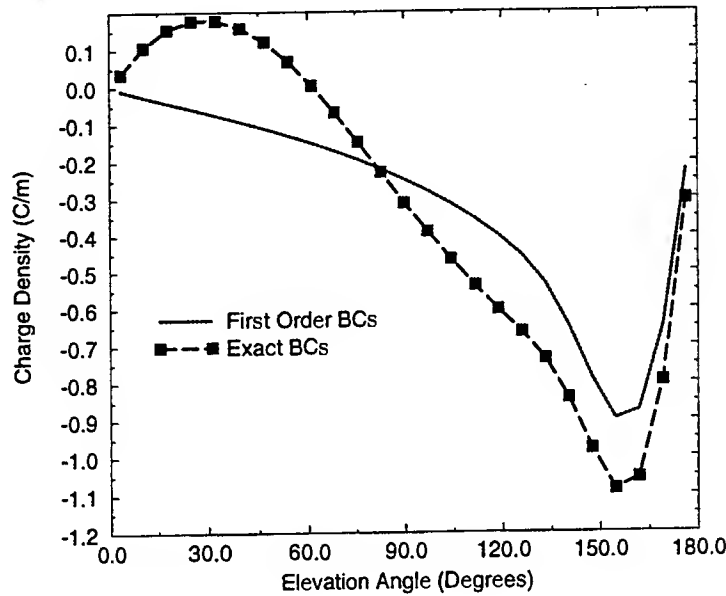


Figure 3: Charge density as a function of elevation angle after 450 time steps;  $\phi = 0$ .

4 and 5, which show the RCS when  $\phi = 0$  and  $\phi = 90^\circ$ , respectively; also shown on this plot is the theoretical data obtained from the Mie series. All data corresponds to a sphere whose electrical radius,  $ka$ , is 6.0. To obtain this data, Maxwell's equations are cast in terms of scattered field quantities. To generate frequency-domain information from the time-domain data, a processing algorithm is required that computes the running Fourier transform. The RCS data is obtained by invoking Schelkunoff's principle of equivalent sources. Since the finite-volume formulation requires the knowledge of surface cell areas in the flux evaluations, the computation of the Schelkunoff surface integrals is reduced to a sum of these cell areas weighted by the integrand at the cell center.

Obviously, all RCS solutions tend to follow the general undulations of the theoretical data and capture the correct value for the forward-scattered RCS. Unfortunately, the scheme that uses the exact boundary condition implementation appears to deviate more from the theoretical solution in the backscattered region.

Numerical simulation has proven that the derived Euler equations indeed predict the surface charge and current on a perfectly conducting surface. However, after numerous simulations, we have found that the algorithm is subject to late-time instabilities. For example, when the same simulation that was used to create the previous RCS data is increased to 2000 time steps, the data no longer showed any trend towards convergence. The reasons for this remain undiscovered. Two possibilities have been identified: 1) The

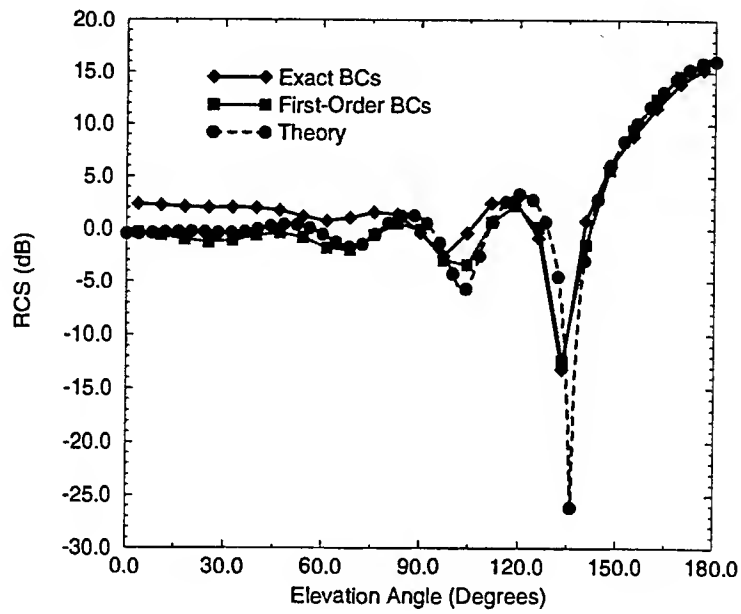


Figure 4: RCS of a sphere when  $ka = 6$  and  $\phi = 0$ .

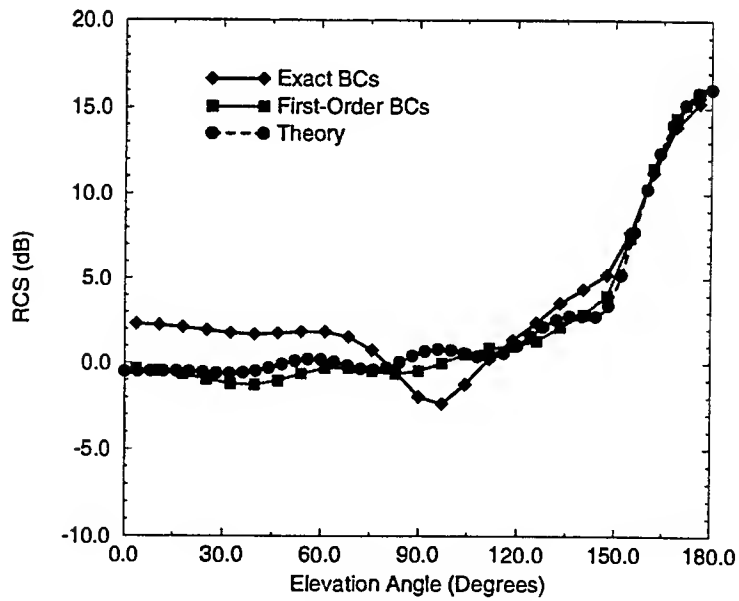


Figure 5: RCS of a sphere when  $ka = 6$  and  $\phi = 90^\circ$ .

source term in Euler's equations is not truly a source term but a field term that is computed from Maxwell's equations. Hence, the way this term is computed can effect stability. For the present implementation, one-sided differencing that is second-order accurate is used. 2) Currently, the boundary conditions are applied at the cell centers rather than at the cell walls. It has been suggested that such an approach reduces the accuracy of the scheme, thereby inducing possible instabilities.

## 5 Conclusion

A formal treatment for extending Maxwell's equations to the boundary has been given. Based upon preliminary numerical experimentations, these new equations, identified as Euler's equations, indeed predict the space-time profile of the surface charge and surface current. Unfortunately, when cast in terms of a finite volume procedure, various simulations have shown that the data may become unstable for a highly resolved grid or for late-time simulations. The reasons for this instability is the subject of future work.

Although future work should include an analytical treatment of the Neumann stability properties, the treatment is not trivial. First, most stability analyses start by considering a one-dimensional domain with periodic boundary conditions. For our situation, the domain is not periodic but is terminated by a perfect conductor. Such a termination greatly complicates the analysis. Moreover, a one-dimensional plane wave is by definition transverse to the perfect conductor - hence, it cannot induce any surface charge. Thus, Euler's equations are reduced to a single equation for which the time-rate-of-change of the current density equals the normal derivative of the tangential electric field.

Other future work includes 1) the incorporation of the boundary conditions at flux walls rather than at cell centers and 2) simplifying the problem such that the scatterer's geometry coincides with a Cartesian grid. This latter study would identify whether or not the curvilinear coordinate transformation has any impact on numerical instability.

Finally, in this project the finite-volume code that incorporates either the first-order boundary condition or the Euler-type boundary condition is streamlined for robust performance. The code is vectorized to accommodate the vector nature of the Cray machines. In addition, the code is customized for low memory usage. For example, for a grid size of (50,55,104) and for a simulation time of 4000 time steps, the code requires only 15 Megawords of memory and 2.7 cpu hours on a Cray YMP; on Cray C90, 1.3 cpu hours are needed. This effort in improving the code performance will benefit other DoD projects currently being managed by the PI.

## References

- [1] Yee, K. S., "Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media," *IEEE Trans. Ant. Propagat.*, vol. 14, pp. 302-307, 1966.
- [2] Taflove, A., and M. E. Morris, "Numerical solution of steady-state electromagnetic scattering problems using the time-dependent Maxwell's equations," *IEEE Trans. Microwave Theory Tech.*, vol. 23, pp. 623-630, 1975.
- [3] Madsen, N. K., "Divergence preserving discrete surface integral methods for Maxwell's curl equations using non-orthogonal unstructured grids," *Research Institute for Advanced Computer Science Tech. Rep. 92.04*, NASA Ames Research Center, 1992.
- [4] Mahadevan, K., and R. Mittra, "Radar cross section computation of inhomogeneous scatterers using edge-based finite element methods in the frequency and time domains," *Radio Science*, vol. 28, pp. 1181-1193, 1993.
- [5] Shankar, V., A. H. Mohammadian and W. F. Hall, "A time-domain, finite-volume treatment for the Maxwell equations," *Electromagnetics*, vol. 10, pp. 127-145, 1990.
- [6] Shankar, V., W. F. Hall, A. Mohammadian and C. Rowell, "Algorithmic aspects and supercomputing trends in computational electromagnetics," *AIAA 31st Aerospace Sciences Meeting & Exhibit*, Reno, NV, AIAA 93-0367, 1993.
- [7] Shankar, V., W. F. Hall and S. Palaniswamy, "Accuracy issues in time-domain CEM using structured/unstructured grid formulations," *11th Annual Review of Progress in Applied Computational Electromagnetics*, Monterey, CA, pp. 1185-1192, 1995.
- [8] Shang, J. S. and D. Gaitonde, "Characteristic-based, time-dependent Maxwell equations solvers on a general curvilinear frame," *AIAA Journal*, vol. 33, pp. 491-498, 1995.
- [9] Shang, J. S., and D. Gaitonde, "Scattered electromagnetic field of a re-entry vehicle," *J. Spacecraft & Rockets*, vol. 32, pp. 294-301, 1995.
- [10] Shang, J. S., "Characteristic-based algorithms for solving the Maxwell equations in the time-domain," *IEEE Ant. Propagat. Mag.*, vol. 37, pp. 15-25, 1995.



- [11] Anderson, W. K., J. L. Thomas and B. van Leer, "A comparison of finite volume flux vector splittings for the Euler equations," *AIAA 23rd Aerospace Sciences Meeting*, Reno, NV, AIAA 85-0122, 1985.
- [12] Steger, J. L., and R. F. Warming, "Flux vector splitting of the inviscid gasdynamic equations with application to finite-difference methods," *J. Comp. Phys.*, vol. 40, pp. 263-293, 1981.
- [13] Weber, Y. S., "Investigations on the properties of a finite-volume, time-domain method for computational electromagnetics," *26th AIAA Plasma Dynamics and Lasers Conf.*, San Diego, CA, AIAA 95-1964, 1995.

# **PD-Eigenstructure Assignment Control for Multivariable Nonlinear Tracking and Decoupling**

J. Jim Zhu  
Assistant Professor  
Department of Electrical and Computer Engineering

Louisiana State University  
Baton Rouge, LA 70803  
Email: zhu@rsip.lsu.edu

Final Report for:  
Summer Research Extension Program  
Wright Laboratory

Sponsored by:  
Air Force Office of Scientific Research  
Bolling Air Force Base, DC

and

Wright Laboratory

December 1996

# **PD-Eigenstructure Assignment Control for Multivariable Nonlinear Tracking and Decoupling**

J. Jim Zhu

Department of Electrical and Computer Engineering  
Louisiana State University  
Baton Rouge, LA 70803

## **Abstract**

In this research a systematic design procedure for multivariable nonlinear tracking and output/state decoupling control is developed by way of linearization along a nominal trajectory. The resulting linear time-varying (LTV) tracking error dynamics are then stabilized and decoupled using PD-eigenstructure assignment in a way similar to the eigenstructure assignment design for LTI systems. Main Accomplishments of this research include: (i) extension of the PD-spectrum and PD-eigenvector concepts for scalar polynomial differential operator to vector differential polynomial operators, (ii) extension of Silverman-Wolovich (S-W) transformation to the entire class of uniformly completely controllable (u.c.c.) multivariable (MV) LTV systems, (iii) PD-eigenvector based criteria for uniform controllability and observability of MV LTV systems, (iv) stabilization of u.c.c. MV LTV systems by PD-spectrum assignment, (v) output/state decoupling of u.c.c. MV LTV systems by PD-eigenstructure assignment, and (vi) A BTT autopilot with decoupled roll-yaw dynamics using the PD-eigenstructure assignment control. Due to the time constraint, implementation and simulation results are not available at the present. Further research is needed to address the complexity of the implementation, and to validation the theory and design procedure by simulations. A multiobjective PD-eigenstructure assignment concept is also proposed to address an array of challenges posed by modern missile technology. Additional applications of the results of this research can be found in aircraft flight control, spacecraft altitude control, vibration control, robotics and the like.

# PD-Eigenstructure Assignment Control for Multivariable Nonlinear Tracking and Decoupling

J. Jim Zhu

## 1. Introduction

Due to the ever stringent performance requirements and inherently nonlinear, time-varying and highly coupled aerodynamics [1], modern missile control, *e.g.* the bank-to-turn (BTT) missile autopilot, poses a challenge to control design that has not been successfully met [2]. Despite its well-know limitations, gain scheduling (GS) appears to be the focus of the current prominent research efforts [3], [4], [5]. Because of the lack of mathematical theory for time-varying dynamics, the GS approach treats the *nonlinear, time-varying* missile dynamics as *time-invariant* dynamics linearized at *discrete operating states*. An autopilot is then comprised of a series of *linear time-invariant (LTI) controllers* scheduled for these *frozen operating states and frozen time dynamics* using various LTI control design techniques, such as eigenstructure assignment for guidance command tracking and roll-yaw decoupling [3].

Scheduling of frozen-time, frozen-state controllers for *fast* time-varying dynamics is known to be mathematically fallacious, and practically hazardous [4]. Recent research efforts have been directed towards applying robust control techniques to extend the stability margin of GS controllers [3], [5]. While the stability margin at each frozen-state is greatly improved with modern robust controllers, it does not seem to benefit stability margin of the overall system proportionally. Indeed, failures have been reported in these recent attempts [5; pp. 14-15]. It appears that the GS control technique has been stretched to its limit in coping with fast time-varying dynamics.

In addition to the time-varying stabilization problem, guidance command tracking with a nonlinear, time-varying airframe poses other challenges in missile autopilot design. In particular,

normal acceleration tracking with a tail fin controlled missile is known to be nonminimum phase, and the roll-yaw aerodynamics are highly coupled for a BTT missile. Actuator saturation and suppression of unmodeled structural modes are other practical issues that have to be effectively dealt with by an autopilot. These problems become even more difficult to tackle in the presence of fast time-varying system parameters.

In this research, a recently developed differential-algebraic spectral theory for linear time-varying (LTV) systems is applied to nonlinear tracking control, such as the BTT missile autopilot design problem. Specifically, the notions of Parallel D-eigenvalues (PD-eigenvalues) and Parallel D-eigenvectors (PD-eigenvectors) [6] are used to extend existing results [3] on robust GS eigenstructure assignment command tracking and roll-yaw decoupling controllers. The differential-algebraic spectral theory treats time-varying dynamics as such, therefore it would succeed where the conventional frozen-time, frozen-state GS techniques fail.

Recently, some encouraging preliminary results on a pitch autopilot design based on this differential-algebraic spectral theory have been obtained [7], [8]. The design approach is by linearization of the nonlinear airframe along a nominal (command) trajectory to obtain LTV tracking error dynamics. This design approach poses two technical challenges: (i) implementation of the inverse of the nonlinear airframe dynamics to generate the nominal control, and (ii) exponential stabilization of the LTV tracking error dynamics to achieve guidance command tracking. The first problem can be solved by employing a (dynamic) neural network, which is not to be discussed here (*cf.* [8] for more information). The tracking error stabilization is achieved by assigning the extended-mean (EM) of the closed-loop PD-eigenvalues of the LTV error dynamics to the left-half-plane (LHP) of the complex numbers  $\mathbb{C}$ , in a way very similar to the LTI eigenvalue assignment.

Simulation studies reported in [7], [8] showed that the autopilot was capable of angle-of-attack (AOA) and normal acceleration (NA) tracking of various command trajectories throughout the entire flight envelope without explicit scheduling of any controller parameters. Figure 1.1 below shows AOA tracking performance of the autopilot for an unrealistically demanding

sinusoidal trajectory in the presence of all combinations of  $\pm 50\%$  variation of the aerodynamic coefficients, demonstrating excellent robustness. It is noted that this surprisingly large parametric stability margin was not specifically designed for, but a consequence of the trajectory linearization and the correct stability criterion based on PD-eigenvalues for LTV systems, as opposed to the pointwise linearization and frozen-time stability criterion typical of the GS controllers.

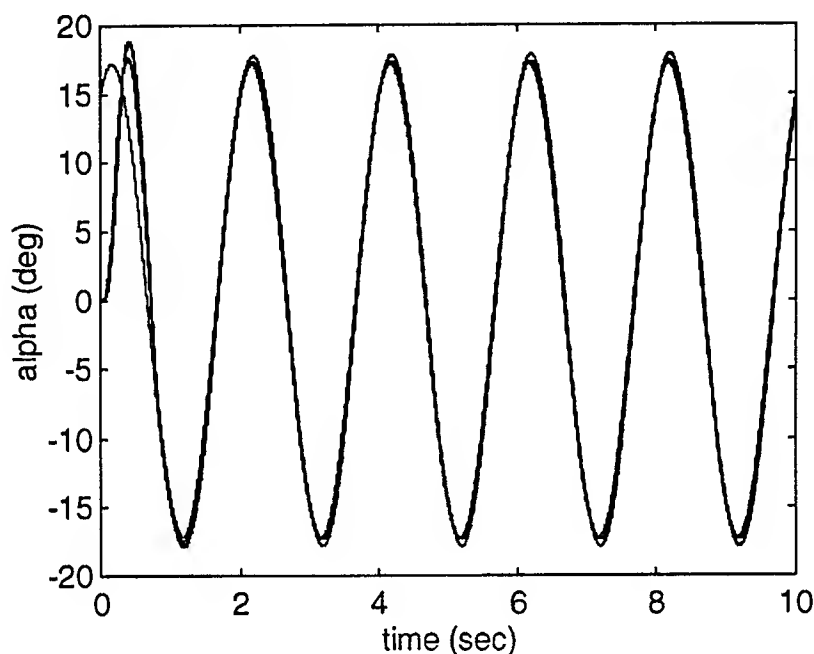


Figure 1.1 Tracking Performance and Robustness;  
with  $\pm 50\%$  Variation on Aerodynamic Coefficients  $C_m$ ,  $C_n$

The present research is a continuation of [7] and a LTV extension of [3]. It is believed to be the first attempt at decoupling of time-varying dynamics using differential-algebraic approach. The design method of [7] is based on the differential-algebraic spectral theory for the class of  $n$ th-order scalar LTV dynamical systems of the form:

$$y^{(n)} + \alpha_n(t)y^{(n-1)} + \cdots + \alpha_2(t)\dot{y} + \alpha_1(t)y = 0 \quad (1.1)$$

$$y^{(k)}(t_0) = y_{k0}, \quad k = 0, 1, \dots, n-1$$

which can be conveniently represented as  $\mathcal{D}_\alpha\{y\} = 0$  using the scalar polynomial differential operator (SPDO)

$$\mathcal{D}_\alpha = \delta^n + \alpha_n(t)\delta^{n-1} + \cdots + \alpha_2(t)\delta + \alpha_1(t) \quad (1.2)$$

where  $\delta = d/dt$  is the derivative operator. It is well-known that the subclass of LTI systems (1), where  $\alpha_k(t) \equiv \alpha_k$ , enjoys an *algebraic* spectral theory that facilitates analytical solutions, precise stability criteria, frequency domain analysis and synthesis, and (robust) stabilization control design techniques. However, as is also well-known, this (time-invariant) algebraic spectral theory does not carry over, in general, to the time-varying case.

The *differential-algebraic* spectral theory for LTV dynamic systems (1) is based on a classical result of Floquet (1879) on the factorization of SPDO [9], [10]

$$\mathcal{D}_\alpha = (\delta - \lambda_n(t)) \cdots (\delta - \lambda_2(t))(\delta - \lambda_1(t)) \quad (1.3)$$

In the differential-algebraic spectral theory, a collection  $\{\lambda_k(t)\}_{k=1}^n$  satisfying (3) is called a *series D-spectrum* (*SD-spectrum*) for  $\mathcal{D}_\alpha$  and an  $n$ -parameter family  $\{\rho_k(t) = \lambda_{1,k}(t)\}_{k=1}^n$  is called a *parallel D-spectrum* (*PD-spectrum*) for  $\mathcal{D}_\alpha$ , where  $\lambda_{1,k}(t)$  are  $n$  particular solutions for  $\lambda_1(t)$  satisfying some nonlinear independence constraints (*cf.* Definition 2.3 below). The scalar functions  $\lambda_k(t)$  and  $\rho_k(t)$  are called *SD-* and *PD-eigenvalues*, respectively, for (1.1) and (1.2) [6].

Let  $A_c(t)$  be the companion matrix associated with  $\mathcal{D}_\alpha$

$$A_c(t) = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \\ 0 & \cdots & \cdots & 0 & 1 \\ -\alpha_1(t) & -\alpha_2(t) & \cdots & \cdots & -\alpha_n(t) \end{bmatrix} \quad (1.4)$$

The matrix

$$\Gamma(t) = \begin{bmatrix} \lambda_1(t) & 1 & 0 & \cdots & 0 \\ 0 & \lambda_2(t) & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & 1 \\ 0 & \cdots & \cdots & 0 & \lambda_n(t) \end{bmatrix} \quad (1.5)$$

is called a *Series Spectral canonical form* (SS canonical form) for  $\mathcal{D}_\alpha$  and  $A_c(t)$ . The diagonal matrix

$$\mathcal{R}(t) = \text{diag} [\rho_1(t), \rho_2(t), \dots, \rho_n(t)] \quad (1.6)$$

is called a *Parallel Spectral canonical form* (PS canonical form) for  $\mathcal{D}_\alpha$  and  $A_c(t)$ . Associated with every PS canonical matrix, there is a *canonical modal matrix* given by

$$V(\rho_1, \rho_2, \dots, \rho_n) = \begin{bmatrix} 1 & 1 & \dots & 1 \\ \mathcal{Q}_{\rho_1}(1) & \mathcal{Q}_{\rho_2}(1) & \dots & \mathcal{Q}_{\rho_n}(1) \\ \mathcal{Q}_{\rho_1}^2(1) & \mathcal{Q}_{\rho_2}^2(1) & \dots & \mathcal{Q}_{\rho_n}^2(1) \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{Q}_{\rho_1}^{n-1}(1) & \dots & \dots & \mathcal{Q}_{\rho_n}^{n-1}(1) \end{bmatrix} \quad (1.7)$$

where  $\mathcal{Q}_{\rho_i} = (\delta + \rho_i)$ ,  $\mathcal{Q}_{\rho_i}^k = \mathcal{Q}_{\rho_i} \mathcal{Q}_{\rho_i}^{k-1}$ . It is noted that the column vectors  $v_i(t)$  of  $V(t)$  satisfy

$$A_c(t)v_i(t) - \rho_i(t)v_i(t) = \dot{v}_i(t) \quad (1.8)$$

and the row vectors  $u_i^T(t)$  of  $U(t) = V^{-1}(t)$  satisfy

$$u_i^T(t)A_c(t) - \rho_i(t)u_i^T(t) = -\dot{u}_i^T(t) \quad (1.9)$$

Thus,  $v_i(t)$  and  $u_i^T(t)$  have been called *column PD-eigenvectors* and *row PD-eigenvectors*, respectively, of  $\mathcal{D}_\alpha$  and  $A_c$  associated with  $\rho_i(t)$ . SD-eigenvectors can be defined similarly [6].

In this research, the concepts of SD-, PD-eigenvalues, and SD-, PD-eigenvectors for scalar LTV systems will first be extended to the class of  $n$ -dimensional,  $l$ -input,  $m$ -output multivariable (MV) LTV systems of the form

$$\begin{aligned} \dot{x} &= A(t)x + B(t)u \\ y &= C(t)x + D(t)u \end{aligned} \quad (1.10)$$

The following basic assumptions are made throughout this report:

- (a) The parameter matrices  $A(t)$ ,  $B(t)$ ,  $C(t)$  are sufficiently smooth functions of time which are bounded, and have bounded, continuous derivatives up to  $(n - 1)$  times.
- (b) The matrix  $D(t)$  is a bounded continuous function of  $t$ .
- (c) For all allowable parameter values,  $\text{rank } B(t) \equiv l$  and  $C(t) \equiv m$ .



- (d) The pair  $\{A(t), B(t)\}$  is uniformly completely controllable (u.c.c.) for all allowable parameter values.
- (e) The pair  $\{A(t), C(t)\}$  is uniformly completely observable (u.c.o.) for all allowable parameter values, if a state observer is needed.

The extension is by way of Lyapunov (coordinate) transformations as outlined in [11; Chapter 5], which allows the Silverman-Wolovich (S-W) transformation [12], [13] to be employed for the assignment of PD-eigenstructures. As an incidental result, a limitation of the S-W transformation is removed so that it is now applicable to the entire class of u.c.c. MV LTV systems. PD-eigenvector based criteria are then obtained for uniform controllability and uniform observability of MV LTV systems, which are natural LTV counterparts of the well known Popov-Belevitch-Hautus eigenvector tests for controllability and observability of LTI systems [14]. Together these results allow stabilization and output/state decoupling of u.c.c. MV LTV systems by PD-eigenstructure assignment, which will be applied to a BTT missile autopilot design using an approach parallel to that of [3].

The extension of the differential-algebraic spectral theory is presented in Section 2. Section 3 presents the extended S-W transformation. The PD-eigenvector criteria for uniform controllability and uniform observability are given in Section 4. In section 5, the extended mean stability criterion for scalar LTV systems is generalized to MV LTV systems, and stabilization of u.c.c. MV LTV systems by PD-eigenvalue assignment is discussed. Section 6 is devoted to output/state decoupling by PD-eigenstructure assignment. The main results of Sections 2-6 are then applied in Section 7 to a BTT missile autopilot design. The report is concluded with Section 8 containing a summary of the main results and suggestions for further studies along this direction.

## 2. PD-eigenstructure for VPDO

Let  $\mathbb{K}$  be the differential ring of  $C^\infty$  functions on  $[0, \infty)$ . Let  $\mathbb{K}^n$  be the  $n$ -dimensional differential module of  $n$ -vectors  $v(t) = \text{col}[v_i(t)]$ , and  $\mathbb{K}^{n \times n}$  be the differential module of  $n \times n$

matrices  $A(t) = [a_{ij}(t)]$ , with entries  $v_i$  and  $a_{ij}$  from  $\mathbb{K}$ . The following two  $n$ -dimensional, first-order, mutually adjoint vector polynomial differential operators (VPDO)

$$\mathcal{P}_A = \delta - A(t) = \mathcal{Q}_{(-A^T)} \quad (2.1)$$

and

$$\mathcal{Q}_A = \delta + A^T(t) = \mathcal{P}_{(-A^T)} \quad (2.2)$$

play an instrumental role in the development of a differential-algebraic spectral theory for both LTI and LTV systems. For instance, a MV LTV system (1.10) can be represented by

$$\begin{aligned} \mathcal{P}_A x &= B(t)u \\ y &= C(t)x + D(t)u \end{aligned} \quad (2.3)$$

Moreover, if we define the inverse VPDO  $\mathcal{P}_A^{-1} = [\delta I - A(t)]^{-1}$  as the integral operator such that  $\mathcal{P}_A^{-1}\mathcal{P}_A = I$ , where  $I$  is the identity operator, then the output  $y(t)$  with zero initial conditions can be conveniently represented by

$$y(t) = [C(t)[\delta I - A(t)]^{-1}B(t) + D(t)]u(t) \quad (2.4)$$

In the sequel, we shall adopt the convention that  $\mathcal{P}_A^0 = I$ , the identity operator, and  $\mathcal{P}_A^k = \mathcal{P}_A \mathcal{P}_A^{k-1}$ . The same applies to  $\mathcal{Q}_A$ . Although the VPDOs  $\mathcal{P}_A$  and  $\mathcal{Q}_A$  are defined for  $n$ -vectors  $v \in \mathbb{K}^n$ , we will also use them on matrices  $M \in \mathbb{K}^{n \times r}$  in a columnwise fashion. For  $n = 1$ ,  $A(t)$  becomes a scalar function, say  $a(t)$ , and the VPDOs  $\mathcal{P}_A$  and  $\mathcal{Q}_A$  become SPDOs denoted by  $\mathcal{P}_a$  and  $\mathcal{Q}_a$ , respectively.

A set of  $n$  vectors  $\{l_k\}_{k=1}^n$ ,  $l_k \in \mathbb{K}^n$  is called a *uniform basis* for  $\mathbb{R}^n$  if  $|\det L(t)| \geq \delta$  for some  $\delta > 0$ , for all  $t \geq 0$ , where  $L(t) = [l_1(t) \mid l_2(t) \mid \cdots \mid l_n(t)]$ . A uniform basis  $\{l_k\}_{k=1}^n$  satisfying  $\|L(t)\| \leq M$  and  $\|\dot{L}(t)\| \leq M$  for some  $M < \infty$ , for all  $t \geq 0$ , is called a *Lyapunov basis* for  $\mathbb{R}^n$ . The matrix  $L(t)$  is called the *coordinate transformation matrix* from the standard basis  $\{e_k\}_{k=1}^n$  for  $\mathbb{R}^n$  to the basis  $\{l_k\}_{k=1}^n$ , where  $e_k$  denotes the  $k$ th column vector of the identity matrix  $I$ . The matrix  $L(t)$  is called a *Lyapunov transformation matrix* if  $\{l_k\}_{k=1}^n$  is a Lyapunov

basis. With these terminology, the PD-eigenvalue and PD-eigenvector concepts for SPDO are extended to VPDO in the following definition.

**Definition 2.1.**

(a) A continuously differentiable scalar function  $\rho(t)$  is called a *PD-eigenvalue* of an  $n$ -dimensional VPDO  $\mathcal{P}_A$  if there exists a Lyapunov transformation matrix  $L(t)$  such that the vector

$$\mathbf{p}(t) = L(t)\mathbf{p}_0(\rho(t)) \quad (2.5)$$

where

$$\mathbf{p}_0(\rho(t)) = \begin{bmatrix} 1 \\ \mathcal{Q}_\rho(1) \\ \vdots \\ \mathcal{Q}_\rho^{n-1}(1) \end{bmatrix} \quad (2.6)$$

satisfies  $\mathcal{P}_{[A-\rho I]}\mathbf{p} = 0$ , or what is the same

$$\dot{\mathbf{p}}(t) = [A(t) - \rho(t)I]\mathbf{p}(t) \quad (2.7)$$

The vector  $\mathbf{p}(t)$  is then called a *PD-eigenvector* of  $\mathcal{P}_A$  associated with  $\rho(t)$ .

(b) Let  $\rho(t)$  be a PD-eigenvalue of  $\mathcal{P}_A$ . A vector  $\mathbf{q}(t)$  satisfying  $\mathcal{Q}_{[A-\rho I]}\mathbf{q} = 0$ , or what is the same

$$\dot{\mathbf{q}}(t) = -[A(t) - \rho(t)I]^T\mathbf{q}(t) \quad (2.8)$$

is called an *adjoint PD-eigenvector* of  $\mathcal{P}_A$  associated with  $\rho(t)$ .

(c) Let  $\mathbf{p}(t)$  and  $\mathbf{q}(t)$  be a PD-eigenvector and an adjoint PD-eigenvector for  $\mathcal{P}_A$  associated with a PD-eigenvalue  $\rho(t)$ . Then  $\rho(t)$  is called a *PD-eigenvalue* of  $A(t)$ . The vectors  $\mathbf{p}(t)$  and  $\mathbf{q}^T(t)$  are called a *column PD-eigenvector* and a *row PD-eigenvector*, respectively, of  $A(t)$  associated with  $\rho(t)$ .

**Remarks.**

1. Let  $\beta(t) = \{l_1(t), l_2(t), \dots, l_n(t)\}$ , where  $l_k(t)$  are the  $k$ th column vector of  $L(t)$  in (a) of Definition 2.1. Then  $\beta(t)$  constitutes a Lyapunov basis with respect to which the PD-eigenvector  $p(t)$  of  $\mathcal{P}_A$  can be represented by  $p_0(\rho(t))$ .
2. It follows from (7) and (8) that if  $p(t)$  and  $q(t)$  are column and row PD-eigenvectors, respectively, of  $A(t)$  associated with a PD-eigenvalue  $p(t)$ , then  $\rho(t)$  is also a PD-eigenvalue of  $-A^T(t)$  with an associated column PD-eigenvector  $-q(t)$  and a row PD-eigenvector  $-p^T(t)$ .

The following definition introduces the notions of a differentially distinct set. This notion is subsequently used to define the concept of a PD-spectrum of a VPDO.

**Definition 2.2.**

Let  $\{\rho_i(t)\}_{i=1}^k$  be a set of  $k$  PD-eigenvalues of  $A(t)$ . The set is said to be *differentially distinct* if the associated set of column PD-eigenvectors  $\{p_i(t)\}_{i=1}^k$  is linearly independent.

**Remark.**

Being in a set,  $\rho_i(t)$  are distinct in the sense that  $\rho_i(t) \neq \rho_j(t)$ . However, they are not necessarily differentially distinct. Consider, for example, the set  $\{\rho_i(t)\}_{i=1}^3 = \{-\frac{1}{2}, \frac{1}{2}, \frac{e^t-1}{2(e^t+1)}\}$  for  $A = \text{comp}[1, 0.25, -4]$ . The associated column PD-eigenvectors are

$$p_1 = \begin{bmatrix} 1 \\ -\frac{1}{2} \\ \frac{1}{4} \end{bmatrix}, \quad p_2 = \begin{bmatrix} 1 \\ \frac{1}{2} \\ \frac{1}{4} \end{bmatrix}, \quad p_3 = \begin{bmatrix} 1 \\ \frac{e^t-1}{2(e^t+1)} \\ \frac{1}{4} \end{bmatrix}$$

which are clearly linearly dependent.

**Definition 2.3.**

(a) A differentially distinct set of  $n$  PD-eigenvalues  $\{\rho_i(t)\}_{i=1}^n$  for an  $n$ -dimensional VPDO  $\mathcal{P}_A$  is called a *PD-spectrum* for  $\mathcal{P}_A$ , and for the associated  $n \times n$  matrix  $A(t)$ .

(b) A PD-spectrum  $\{\rho_i(t)\}_{i=1}^n$  for an  $n$ -dimensional VPDO  $\mathcal{P}_A$  together with a set of associated PD-eigenvectors  $\{p_i(t)\}_{i=1}^n$  is called a *PD-eigenstructure* for  $\mathcal{P}_A$ , and for the associated  $n \times n$  matrix  $A(t)$ .

**3. Extension of Silverman-Wolovich Transformation**

Using the VPDO  $\mathcal{P}_A$ , the controllability matrix for MV LTV system (10) can be written as

$$C(t) = [B(t) \mid \mathcal{P}_A B(t) \mid \cdots \mid \mathcal{P}_A^{n-1} B(t)]$$

The pair  $\{A(t), B(t)\}$  is u.c.c. if and only if  $\text{rank } C(t) \equiv n$ . Thus, if  $\{A(t), B(t)\}$  is u.c.c., then for any fixed  $t$ , there exists indices  $n_1(t), n_2(t), \dots, n_l(t)$  with  $\sum_{j=1}^l n_j(t) = n$  such that

$$\text{rank } P(t) = \text{rank } [P_1(t) \mid P_2(t) \mid \cdots \mid P_l(t)] = n$$

where

$$P_j(t) = [p_{j1}(t) \mid p_{j2}(t) \mid \cdots \mid p_{jn_j(t)}(t)]$$

with

$$p_{jk}(t) = \mathcal{P}_A^{k-1} b_j(t)$$

where  $b_j(t)$  is the  $j$ th column vector of  $B(t)$ . The set  $\beta = \{p_{jk}(t)\}$  of the  $n$  column vectors of  $P(t)$  is called a *lexicographic basis* at  $t$  for  $\mathbb{R}^n$  generated by  $\{A(t), B(t)\}$ , and the set of indices  $\{n_j(t)\}$  is called a set of *lexicographic indices*. A u.c.c. pair of  $\{A(t), B(t)\}$  is said to have a *lexicographic Lyapunov basis* if a set of constant lexicographic indices  $\{n_j\}$  can be chosen for all  $t$ .

Let  $\beta$  be a lexicographic Lyapunov basis. Let  $d_k = \sum_{j=1}^k n_j$  with  $d_0 = 0$ , where  $n_j$  are the lexicographic indices for  $\beta$ . Then the generic multi-variable phase variable (MVPV) canonical form  $\{A_p(t), B_p(t)\}$  associated with  $\beta$  is given by

$$A_p(t) = \begin{bmatrix} A_{11}(t) & A_{12}(t) & \cdots & A_{1l}(t) \\ A_{21}(t) & A_{22}(t) & \cdots & A_{2l}(t) \\ \vdots & \vdots & \ddots & \vdots \\ A_{l1}(t) & A_{l2}(t) & \cdots & A_{ll}(t) \end{bmatrix} \quad (3.1)$$

$$B_p(t) = [b_{p1} \mid b_{p2} \mid \cdots \mid b_{pl}] \quad (3.2)$$

where

$$A_{ii}(t) = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ \alpha_{i,d_{i-1}+1} & \alpha_{i,d_{i-1}+2} & \alpha_{i,d_{i-1}+3} & \cdots & \alpha_{i,d_i} \end{bmatrix}$$

$$A_{ik}(t) = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \\ \alpha_{i,d_{k-1}+1} & \alpha_{i,d_{k-1}+2} & \alpha_{i,d_{k-1}+3} & \cdots & \alpha_{i,d_k} \end{bmatrix}$$

and  $b_{pk} = e_{d_k}$ , where  $e_i$  is the  $i$ th standard basis for  $\mathbb{R}^n$ .

If a pair  $\{A(t), B(t)\}$  is u.c.c. with a lexicographic Lyapunov basis, then it can be reduced to the MVPV canonical form  $\{A_p(t), B_p(t)\}$  by a Lyapunov transformation. Control synthesis can then be done in the MVPV canonical form by state feedback. The Lyapunov transformation matrix can be obtained by an algorithm developed by Silverman (1965) for single input LTV systems, and by Wolovich [13] (1968), (see also [15] by Seal and Stubberud, 1969), for multi-input LTV systems.

The main result of this section extends the Silverman-Wolovich (S-W) transformation to the entire class of u.c.c. MV LTV systems. For u.c.c. systems without a lexicographic Lyapunov basis, an input coordinate transformation is applied so that a lexicographic Lyapunov basis is obtained for the transformed inputs. The S-W transformation algorithm is modified to avoid the inversion of the lexicographic Lyapunov basis matrix  $P(t)$  by orthonormalizing the lexicographic

Lyapunov basis. The VPDO and the adjoint VPDO notations are used to simplify the representation of the algorithm.

**Theorem 3.1.**

*Let the MV LTV system (1.10) be uniformly completely controllable. Then by a state coordinate transformation*

$$\mathbf{x}(t) = \mathbf{L}(t)\mathbf{z}(t)$$

*and an input coordinate transformation*

$$\mathbf{u}(t) = \mathbf{T}(t)\mathbf{v}(t)$$

*the MV LTV system (1.10) can be reduced to the MVPV canonical form*

$$\begin{aligned}\dot{\mathbf{z}} &= \mathbf{A}_p(t)\mathbf{z} + \mathbf{B}_p(t)\mathbf{v} \\ \mathbf{y} &= \mathbf{C}_p(t)\mathbf{z} + \mathbf{D}_p(t)\mathbf{v}\end{aligned}\tag{3.3}$$

*with lexicographic indices  $n_1, n_2, \dots, n_l$ , where*

$$\begin{aligned}\mathbf{A}_p(t) &= \mathbf{L}^{-1}(t)\mathcal{P}_A\mathbf{L}(t) \\ \mathbf{B}_p(t) &= \mathbf{L}^{-1}(t)\mathbf{B}(t)\mathbf{T}(t)\end{aligned}$$

*are of the form (3.1) and (3.2), and*

$$\begin{aligned}\mathbf{C}_p(t) &= \mathbf{C}(t)\mathbf{L}(t) \\ \mathbf{D}_p(t) &= \mathbf{D}(t)\mathbf{T}^{-1}(t)\end{aligned}$$

**Proof (Outline).** For a u.c.c. pair  $\{\mathbf{A}(t), \mathbf{B}(t)\}$  without a lexicographic Lyapunov basis, first apply the elementary column operations of the third kind on the controllability matrix  $\mathcal{C}(t)$ , namely adding to a column another column multiplied by a constant, to obtain a Lyapunov basis. These elementary operations can be facilitated by a constant coordinate transformation in the input space

$$\mathbf{u}(t) = \mathbf{T}_1(t)\mathbf{v}_1(t)$$

Then in the transformed coordinates the pair  $\{A(t), B_1(t)\}$  will have a lexicographic Lyapunov basis, where

$$B_1(t) = B(t)T_1(t) = [\bar{b}_1(t) \quad \bar{b}_2(t) \quad \dots \quad \bar{b}_l(t)]$$

The MVPV canonical form  $\{A_p(t), B_p(t)\}$  can now be obtained from  $\{A(t), B_1(t)\}$  by a state coordinate transformation  $x(t) = L(t)z(t)$  and an input coordinate transformation  $v_1(t) = T_2(t)v(t)$  as follows.

Let  $\beta = \{p_{jk}(t) = \mathcal{P}_A^{k-1}\bar{b}_j(t)\}$  be an *orthonormal* lexicographic Lyapunov basis for  $\{A(t), B_1(t)\}$  with lexicographic indices  $n_1, n_2, \dots, n_l$ . Then the state coordinate transformation matrix  $L(t)$  is given in terms of  $R(t) = L^{-1}(t)$  by

$$R(t) = \begin{bmatrix} \frac{R_1(t)}{R_2(t)} \\ \vdots \\ \frac{R_l(t)}{R_l(t)} \end{bmatrix}$$

where the  $n_i \times n$  submatrices  $R_i(t)$  are given by

$$R_i(t) = \begin{bmatrix} \frac{r_{i1}^T(t)}{r_{i2}^T(t)} \\ \vdots \\ \frac{r_{i,n_i}^T(t)}{r_{i,n_i}^T(t)} \end{bmatrix}$$

where

$$r_{ik}(t) = Q_A^{k-1} \mathcal{P}_A^{n_i-1} \bar{b}_i(t)$$

If  $\beta$  is not orthonormal, let  $P(t)$  be the matrix associated with  $\beta$  and let  $G(t)$  be the left (row) orthonormalization matrix for  $P(t)$  such that

$$\hat{P}(t) = G(t)P(t)$$

is unitary, and let

$$\hat{A}(t) = G(t)P_A G^{-1}(t)$$



$$\hat{B}(t) = G(t)B(t)$$

Then the State transformation matrix  $L(t)$  can be written as

$$L(t) = \hat{L}(t)G^{-1}(t)$$

where  $\hat{L}(t)$  is obtained from the orthonormalized basis represented by  $\hat{P}$  and the system  $\{\hat{A}(t), \hat{B}(t)\}$  as described above.

The input coordinate transformation matrix  $T_2(t)$  is the right (column) orthonormalization matrix for  $R(t)B_1(t)$ . The overall input coordinate transformation  $T(t)$  is given by  $T(t) = T_2(t)T_1(t)$ .  $\square$

A MAPLE implementation of the Extended S-W transformation has been developed. It will be used in the design and implementation of the eigenstructure assignment control in Sections 5, 6 and 7.

#### 4. A PD-Eigenvector Based Criteria for Uniform Controllability and Observability

In this section PD-eigenvector based criteria for uniform controllability and observability are obtained. These results are important in their own right because they are natural LTV counterparts to the well known Popov-Belevitch-Hautus eigenvector tests for controllability and observability of LTI systems [14]. They will also constitute the basis for output/state decoupling by PD-eigenstructure assignment. The first main result of this section gives a necessary and sufficient condition on the pointwise controllability and observability using PD-eigenvectors.

**Theorem 4.1** (PD-eigenvector criteria for controllability and observability).

*A MV LTV system  $\{A(t), B(t), C(t)\}$  is not completely controllable at  $t_1$  if and only if there is a row PD-eigenvectors  $q^T(t)$  of  $A(t)$  such that  $\gamma(t_1) = q^T(t_1)B(t_1) = 0$  and  $\dot{\gamma}(t_1) = 0$ . It is not completely observable at  $t_1$  if and only if there is a column PD-eigenvector  $p(t)$  of  $A(t)$  such that  $\eta(t_1) = C(t_1)p(t_1) = 0$  and  $\dot{\eta}(t_1) = 0$ .*

**Proof.** The proof will be given for the controllability statement only, as the statement for the observability can be proved by proving controllability for the adjoint system. Let  $\mathcal{C}(t) = [\mathcal{C}_1(t) \mid \mathcal{C}_2(t) \mid \cdots \mid \mathcal{C}_n(t)]$  be the controllability matrix for  $\{A(t), B(t)\}$ , where  $\mathcal{C}_k(t) = \mathcal{P}_A^{k-1} B(t)$ . Suppose that  $q^T(t)$  is a row PD-eigenvector for  $A(t)$  associated with a PD-eigenvalue  $\rho(t)$  such that  $\gamma_1(t_1) = q^T(t_1)B(t_1) = q^T(t_1)\mathcal{C}_1(t_1) = 0$  and  $\dot{\gamma}_1(t_1) = 0$ . Let  $\gamma_k(t_1) = q^T(t_1)\mathcal{C}_k(t_1)$ . We shall first show by induction that  $\gamma_k(t_1) = \rho(t_1)\gamma_{k-1}(t_1) = 0$  and  $\dot{\gamma}_k(t_1) = 0$ ,  $k = 2, 3, \dots, n$ . To this end, note that  $\dot{\gamma}_{k-1}(t_1) = 0$  implies that  $-q^T(t_1)\dot{\mathcal{C}}_{k-1}(t_1) = \dot{q}^T(t_1)\mathcal{C}_{k-1}(t_1)$ . It then follows from the induction hypothesis that

$$\begin{aligned}\gamma_k(t_1) &= q^T(t_1)\mathcal{C}_k(t_1) \\ &= q^T(t_1)\mathcal{P}_A\mathcal{C}_{k-1}(t_1) \\ &= q^T(t_1)A(t_1)\mathcal{C}_{k-1}(t_1) - q^T(t_1)\dot{\mathcal{C}}_{k-1}(t_1) \\ &= [q^T(t_1)A(t_1) + \dot{q}^T(t_1)]\mathcal{C}_{k-1}(t_1) \\ &= \rho(t_1)q^T(t_1)\mathcal{C}_{k-1}(t_1) \\ &= \rho(t_1)\gamma_{k-1}(t_1) \\ &= 0\end{aligned}$$

Moreover

$$\dot{\gamma}_k(t_1) = \dot{\rho}(t_1)\gamma_{k-1}(t_1) + \rho(t_1)\dot{\gamma}_{k-1}(t_1) = 0$$

Consequently,  $q^T(t_1)\mathcal{C}(t_1) = 0$ . Since  $q^T(t_1) \neq 0$ ,  $\text{rank } \mathcal{C}(t_1) < n$ . Thus  $\{A(t), B(t)\}$  is not completely controllable at  $t_1$ .

Conversely, suppose that  $\text{rank } \mathcal{C}(t_1) = r < n$ , so that  $\{A(t), B(t)\}$  is not completely controllable at  $t_1$ . Then there exists a Lyapunov basis  $\Gamma$  with respect to which  $\{A(t), B(t)\}$  has a representation  $\{\hat{A}(t), \hat{B}(t)\}$  such that

$$\hat{A}(t_1) = \left[ \begin{array}{c|c} \hat{A}_{11}(t_1) & \hat{A}_{12}(t_1) \\ \hline 0 & \hat{A}_{22}(t_1) \end{array} \right], \quad \hat{B}(t_1) = \left[ \begin{array}{c} \hat{B}_1(t_1) \\ 0 \end{array} \right]$$

where  $\hat{A}_{11}(t_1) \in \mathbb{R}^{r \times r}$  is completely controllable,  $\hat{A}_{22}(t_1) \in \mathbb{R}^{s \times s}$ ,  $s = n - r$ , is completely uncontrollable, and  $\hat{B}_1(t_1) \in \mathbb{R}^{r \times 1}$  is nonzero. Let  $\rho(t)$  be a PD-eigenvalue for  $\hat{A}_{22}(t)$  with an

associated row PD-eigenvector  $\hat{z}^T(t)$ . Let  $\hat{q}^T(t) = q^T(t)L(t) = [0 \mid \hat{z}^T(t)]$ , where  $L(t)$  is the coordinate transformation matrix from the standard basis to  $\Gamma$ . Clearly  $\gamma(t_1) = q^T(t_1)B(t_1) = \hat{q}^T(t_1)L^{-1}(t)L(t)\hat{B}(t_1) = 0$ , and  $\dot{\gamma}(t_1) = 0$ . It remains to show that  $\rho(t)$  is a PD-eigenvalue for  $A(t)$  and  $q^T(t)$  is a row PD-eigenvector for  $A(t)$ . Clearly,  $q^T(t)$  satisfies

$$\begin{aligned} -\dot{\hat{q}}^T(t) &= [0 \mid -\dot{\hat{z}}^T(t)] \\ &= [0 \mid \hat{z}^T(t) [\hat{A}_{22}(t) - \rho(t)I_s]] \\ &= \hat{q}^T(t)\hat{A}(t) - \rho(t)\hat{q}^T(t) \end{aligned}$$

By Remark 2 to Definition 2.1,  $\hat{z}(t)$  is a column PD-eigenvector for  $-\hat{A}_{22}^T(t)$ . Thus, there exists a Lyapunov basis  $\{\zeta_1(t), \zeta_2(t), \dots, \zeta_n(t)\}$  for  $\mathbb{R}^s$  such that

$$\hat{z}(t) = \sum_{i=1}^s [\mathcal{Q}_\rho^{i-1}(1)] \zeta_i(t)$$

Now construct a Lyapunov basis  $\{\beta_1(t), \beta_2(t), \dots, \beta_n(t)\}$  for  $\mathbb{R}^n$  as follows

$$\begin{aligned} \beta_i(t) &= \begin{bmatrix} 0 \\ \zeta_i(t) \end{bmatrix}, \quad i = 1, 2, \dots, s-1 \\ \beta_s(t) &= \begin{bmatrix} \psi_0(t) \\ \zeta_s(t) \end{bmatrix} \\ \beta_j(t) &= e_{j-s}, \quad j = s+1, \dots, n \end{aligned}$$

where  $e_k$  is the  $k$ th standard basis vector for  $\mathbb{R}^n$  and  $\psi_0(t)$  is chosen by

$$\psi_0(t) = -\sum_{j=s+1}^n [\mathcal{Q}_\rho^{j-1}(1)] e_{j-s}$$

Then

$$\hat{q}(t) = \sum_{i=1}^n [\mathcal{Q}_\rho^{i-1}(1)] \beta_i(t)$$

It then follows from Remark 2 to Definition 2.1 that  $\rho(t)$  is a PD-eigenvalue for  $A(t)$  and  $q^T(t)$  is a row PD-eigenvector for  $A(t)$ .  $\square$

In the following corollary to Theorem 4.1, a MV LTV system  $\{A(t), B(t), C(t)\}$  is said to be *uniformly uncontrollable* if the controllability matrix  $C(t)$  for  $\{A(t), B(t)\}$  satisfies  $\text{rank } C(t) < n, \forall t \geq 0$ . It is said to be *uniformly unobservable* if the observability matrix  $O(t)$  for  $\{A(t), C(t)\}$  satisfies  $\text{rank } O(t) < n, \forall t \geq 0$ .

**Corollary 4.1.**

*A MV LTV system  $\{A(t), B(t), C(t)\}$  is uniformly uncontrollable if and only if  $A(t)$  has a row PD-eigenvector  $q^T(t)$  of  $A(t)$  such that  $q^T(t)B(t) \equiv 0$ . It is uniformly unobservable if and only if  $A(t)$  has a column PD-eigenvector  $p(t)$  such that  $C(t)p(t) \equiv 0$ .  $\square$*

In the following corollary to Theorem 4.1, a MV LTV system  $\{A(t), B(t), C(t)\}$  is said to be *uniformly  $r$ -controllable* if the controllability matrix  $C(t)$  for  $\{A(t), B(t)\}$  satisfies  $\text{rank } C(t) \equiv r$ . It is said to be *uniformly  $r$ -observable* if the observability matrix  $O(t)$  for  $\{A(t), C(t)\}$  satisfies  $\text{rank } O(t) \equiv r$ .

**Corollary 4.2.**

*A MV LTV system  $\{A(t), B(t), C(t)\}$  is uniformly  $r$ -controllable if and only if there exist exactly  $s = n - r$  row PD-eigenvectors  $q_i^T(t)$  such that  $q_i^T(t)B(t) \equiv 0, i = 1, 2, \dots, s$ . It is uniformly  $r$ -observable if and only if there exist exactly  $s = n - r$  column PD-eigenvector  $p_j(t)$  such that  $C(t)p_j(t) \equiv 0, j = 1, 2, \dots, s$ .  $\square$*

The following theorem gives PD-eigenvector based criteria for modal controllability and observability. It sheds some light on how the orientation of a PD-eigenvectors affects the “degree” of controllability and observability. The result on modal observability will be used subsequently in Section 6 to develop the output/state decoupling technique by PD-eigenstructure assignment.

**Theorem 4.2** (Modal controllability and observability).

Let  $\{A(t), B(t), C(t)\}$  be a MV LTV system and let  $b_j(t)$  and  $c_i^T(t)$  be the  $j$ th column vector of  $B(t)$  and the  $i$ th row vector of  $C(t)$ , respectively. If  $q_i^T(t)$  is a row PD-eigenvector of  $A(t)$  associated with the  $i$ th PD-eigenvalue  $\rho_i(t)$  such that  $q_i^T(t)b_j(t) \equiv 0$ , then the associated  $i$ th mode  $\exp \int_0^t \rho_i(\tau) d\tau$  cannot be altered by the  $j$ th input  $u_j(t)$  with the state feedback control law  $u_j(t) = k_j^T(t)x(t)$  for any gain  $k_j^T(t)$ . Similarly, If  $p_j(t)$  is a column PD-eigenvector of  $A(t)$  associated with the  $j$ th PD-eigenvalue  $\rho_j(t)$  such that  $c_i^T(t)p_j(t) \equiv 0$ , then the associated  $j$ th mode  $\exp \int_0^t \rho_j(\tau) d\tau$  is not observable from the  $i$ th output  $y_i(t)$  with a state observer for any observer gain  $h_j(t)$ .

**Proof of Theorem 4.2.** The proof will be given only for the controllability statement. The statement for observability can be proved by applying the proof for controllability to the adjoint system. Let  $\{\rho_i(t)\}_{i=1}^n$  be a PD-spectrum for  $A(t)$  and let  $Q(t) = [q_1(t) \mid q_2(t) \mid \dots \mid q_n(t)]^T$  be the associated row PD-modal matrix consisting of the row PD-eigenvectors  $q_k^T(t)$ . Then the coordinate transformation  $z(t) = Q(t)x(t)$  results in  $\dot{z} = A_z(t)z + B_z(t)u$ , where

$$\begin{aligned} A_z(t) &= Q(t)Q_A Q^{-1}(t) \\ &= \text{diag}[\rho_1(t), \rho_2(t), \dots, \rho_n(t)] \end{aligned}$$

$$B_z(t) = Q(t)B(t)$$

Now suppose that  $q_i^T(t)b_j(t) \equiv 0$  and let  $u_j(t) = k_j^T x(t) = k_j^T Q^{-1}(t)z(t)$ . Then

$$\begin{aligned} \dot{z}_i(t) &= \rho_i(t)z_i(t) + q_i^T(t)b_j^T(t)k_j^T(t)Q^{-1}(t)z(t) \\ &= \rho_i(t)z_i(t) \end{aligned}$$

Clearly, the  $i$ th mode  $z_i(t) = \exp \int_0^t \rho_i(\tau) d\tau$  can not be altered by the  $j$ th input  $u_j(t) = k_j^T x(t)$  for any  $k_j^T(t)$ .  $\square$

## Remarks.

1. Although the proof for the observability statement in Theorem 4.2 is omitted, it is instructive to note that  $y_i(t) = c_i^T(t)x(t)$ . Note let  $P(t) = [p_1(t) \mid p_2(t) \mid \dots \mid p_n(t)]$  be the column PD-modal matrix associated with a PD-spectrum  $\{\rho_j(t)\}_{j=1}^n$ . Then the coordinate transformation  $x(t) = P(t)z(t)$  results in  $\dot{z} = A_z(t)z + B_z(t)u$ ,  $y(t) = C_z(t)z$ , where

$$\begin{aligned} A_z(t) &= P^{-1}(t)P_A P(t) \\ &= \text{diag}[\rho_1(t), \rho_2(t), \dots, \rho_n(t)] \\ B_z(t) &= P^{-1}(t)B(t) \\ C_z(t) &= C(t)P(t) \end{aligned}$$

Thus, the  $i$ th output  $y_i(t)$  is given by

$$y_i(t) = \sum_{j=1}^n c_i^T(t)p_j(t)z_j(t)$$

Clearly, if  $c_i^T(t)p_j(t) \equiv 0$ , the  $j$ th forced mode

$$z_j(t) = z_j(0)e^{\int_0^t \rho_j(\tau) d\tau} + \int_0^t e^{-\int_\tau^t \rho(\sigma) d(\sigma)} \sum_{k=1}^l b_{ik}(\tau)u_k(\tau) d\tau$$

will be absent from  $y_i(t)$ . Consequently, the  $j$ th free mode  $\exp \int_0^t \rho_j(\tau) d\tau$  is not observable from  $y_i(t)$ .

2. The above arguments constitute the basis for the output/state decoupling by PD-eigenstructure assignment. If it is desirable to decouple the  $i$ th output  $y_i(t)$  from the  $j$ th closed-loop mode  $\exp \int_0^t \rho_j(\tau) d\tau$ , assign the  $j$ th component of the closed-loop eigenstructure  $\{\rho_j(t), p_j(t)\}$  such that  $c_i^T(t)p_j(t) \equiv 0$ . To decouple the  $i$ th state from the  $j$ th closed-loop mode, let  $c_i^T(t) = e_i^T$ , where  $e_i$  is the  $i$ th standard basis vector.

## 5. Stabilization by PD-Spectrum Assignment

In this section the PD-spectrum base stability criterion for the scalar LTV system (1.1) is first extended to MV LTV systems (1.0). Then a result is presented for feedback stabilization of MV LTV systems using PD-spectrum assignment by way of S-W transformation. The design

procedure for PD-spectrum assignment is also provided. The necessary and sufficient stability criterion based on a PD-spectrum uses an extended-mean concept as defined below.

**Definition 5.1.**

Let  $\sigma : I \rightarrow \mathbb{R}$  be a locally integrable function on the interval  $I = [T_0, \infty)$ . The extended mean of  $\sigma(t)$  over  $I$  is defined by

$$\begin{aligned} \text{em}_{t_0, t \in I}(\sigma(t)) &= \limsup_{T \rightarrow \infty, t_0 \geq T_0} \frac{1}{T} \int_{t_0}^{t_0+T} \sigma(\tau) d\tau \\ &= \lim_{T \rightarrow \infty} \left[ \sup_{t \geq t_0+T, t_0 \geq T_0} \frac{1}{t-t_0} \int_{t_0}^t \sigma(\tau) d\tau \right] \end{aligned}$$

**Theorem 5.1.**

Let  $\mathcal{P}_A$  be a VPDO having a PD-spectrum  $\{\rho_k(t)\}_{k=1}^n$  with  $|\text{Re } \rho_k(t)| < M$ ,  $t \geq 0$ , for some  $M < \infty$ . Let  $\mathbf{p}_k(t)$  and  $\mathbf{q}_k^T(t)$  be a column PD-eigenvector and a row PD-eigenvector associated with  $\rho_k(t)$  respectively. Then the null solution to the LTV system  $\mathcal{P}_A \mathbf{x} = \mathbf{0}$  is uniformly asymptotically stable for all  $t_0 \geq T_0$  if and only if

(i) there exists a  $0 < c_k \leq \infty$  such that

$$\text{em}_{t_0, t \in I}(\text{Re } \rho_k(t)) = -c_k < 0$$

and moreover,

(ii) there exist  $h_k > 0$  and  $0 < d_k < c_k$  such that

$$\|\mathbf{p}_k(t) \mathbf{q}_k^T(t_0)\| < h_k e^{d_k(t-t_0)}$$

for all  $t \geq t_0 \geq T_0$ .  $\square$

**Remarks.**

1. Condition (ii) is automatically satisfied if the imaginary parts of all PD-eigenvalues are of polynomial order or slower; that is, an integer  $m > 0$  exists such that

$$\lim_{t \rightarrow \infty} \frac{\text{Im} \{\rho_k(t)\}}{t^m} = 0, \quad k = 1, 2, \dots, n$$

In particular, it holds if  $\text{Im} \{\rho_k(t)\}$  are uniformly bounded.

2. If  $\liminf_{t_0, t \in I} (\text{Re} \rho_k(t)) > 0$  for some  $t_0 \geq T_0$ , and  $1 \leq k \leq n$ , then the null solution to  $\mathcal{P}_A \mathbf{x} = 0$  is unstable. However, if  $\liminf_{t_0, t \in I} (\text{Re} \rho_k(t)) = 0$  for some  $t_0 \geq T_0$ , and  $1 \leq k \leq n$ , the null solution may be either stable, asymptotically stable, or unstable, but it cannot be exponentially stable.

The proof of Theorem 5.1 is based on the results for SPDO presented in [16] and the fact that Lyapunov transformations preserve stability. Thus it is omitted here. The following Theorem 5.2 facilitates stabilization of MV LTV systems using PD-spectrum assignment by way of Silverman-Wolovich transformations.

### Theorem 5.2.

*Let  $A(t) = \text{diag}[A_1(t), A_2(t), \dots, A_l(t)]$ , where  $A_i \in \mathbb{K}^{n_i \times n_i}$  are bounded companion matrices. If  $\rho(t)$  is a PD-eigenvalue of  $A_i(t)$  for some  $i \leq l$  with an associated column PD-eigenvector  $\mathbf{p}_i(t) \in \mathbb{K}^{n_i}$ , then it is a PD-eigenvalue for  $A(t)$  with an associated column PD-eigenvector  $\mathbf{p}(t)$  generated from  $\mathbf{p}_i(t)$ .*

**Proof.** Without loss of generality, suppose  $\rho(t)$  is a PD-eigenvalue of  $A_1(t)$  with an associated PD-eigenvector  $\mathbf{p}_1(t)$ . For if  $i \neq 1$ , a constant similarity transformation  $\hat{A}(t) = L^{-1}A(t)L$  will swap  $A_i(t)$  and  $A_1(t)$ , and the following arguments remain valid under similarity transformations. Now to show that  $\rho(t)$  is a PD-eigenvalue for  $A(t)$ , we need to construct a PD-eigenvector  $\mathbf{p}(t)$  from  $\mathbf{p}_1(t)$  satisfying

$$\dot{\mathbf{p}}(t) = [A(t) - \rho(t)I]\mathbf{p}(t) \tag{5.1}$$

and



$$\mathbf{p}(t) = \sum_{j=1}^n [\mathcal{Q}_\rho^{j-1}(1)] \beta_j(t) \quad (5.2)$$

for some Lyapunov basis  $\{\beta_j(t)\}_{j=1}^n$ . Let

$$\mathbf{p}(t) = \begin{bmatrix} \mathbf{p}_1(t) \\ 0 \end{bmatrix} \quad (5.3)$$

Clearly,  $\mathbf{p}(t)$  satisfies (5.1). Since  $\mathbf{p}_1(t)$  is a PD-eigenvector for  $\mathbf{A}_1(t)$ , there exists a Lyapunov basis  $\{\zeta_k(t)\}_{k=1}^{n_1}$  for  $\mathbb{R}^{n_1}$ , such that

$$\mathbf{p}_1(t) = \sum_{k=1}^{n_1} [\mathcal{Q}_\rho^{k-1}(1)] \zeta_k(t)$$

Now construct a Lyapunov basis  $\{\beta_k(t)\}_{k=1}^n$  as follows

$$\begin{aligned} \beta_k(t) &= \begin{bmatrix} \zeta_k(t) \\ 0 \end{bmatrix} & k = 1, 2, \dots, n_1-1 \\ \beta_{n_1}(t) &= \begin{bmatrix} \zeta_{n_1}(t) \\ \psi_0(t) \end{bmatrix} \\ \beta_k(t) &= \mathbf{e}_k & k = n_1+1, \dots, n \end{aligned}$$

where  $\mathbf{e}_k$  is the  $k$ th standard basis vector for  $\mathbb{R}^n$ , and

$$\psi_0(t) = -\sum_{k=n_1}^{n-1} [\mathcal{Q}_\rho^k(1)] \mathbf{e}_k$$

Then it is readily verified that  $\mathbf{p}(t)$  satisfies (5.2).  $\square$

### Remark.

Let  $\mathbf{A}(t)$  be a block upper (lower) triangular matrix. For  $\mathbf{A}(t) \equiv \mathbf{A}$ , it holds that if  $\lambda$  is an eigenvalue of a block matrix  $\mathbf{A}_{ii}(t)$  on the diagonal, then  $\lambda$  is an eigenvalue of  $\mathbf{A}$ . However, this is not granted for a time-varying  $\mathbf{A}(t)$ .

The design procedure for PD-spectrum assignment is presented below, along with guidelines on the selection of closed-loop PD-spectrum.

### PD-Spectrum Assignment Procedure.

1. Transform the MV LTV system (1.10) into MVPV canonical form (3.3) by a state coordinate transformation  $x(t) = L(t)z(t)$  and an input coordinate transformation  $u(t) = T(t)v(t)$  per Theorem (3.1).
2. For each block companion matrix  $A_{ii}(t)$  in the MVPV matrix  $A(t)$ , chose the desired PD-eigenvalues and synthesis the coefficients of the SPDO associated with  $A_{ii}(t)$ . Then design the state feedback control law  $v(t) = K_p(t)z(t)$  to obtain the desired closed-loop dynamics in the MVPV coordinates.
3. The actual control law  $u(t) = K(t)x(t)$  is given by  $K(t) = T(t)K_p(t)L^{-1}(t)$ .

### Remarks.

1. For BIBO stability, exponential stability must be achieved by assigning negative extended mean to all the PD-eigenvalues. To this end, it suffices to keep  $\text{Re}\{\rho_k(t)\} \leq -\epsilon < 0$  for some prescribed  $\epsilon > 0$ .
2. No identical PD-eigenvalues should be assigned within each companion block  $A_{ii}(t)$ . For block size larger than  $2 \times 2$ , ensure that all PD-eigenvalues are *differentially distinct*.
3. If a pair of complex conjugate PD-eigenvalues  $\rho_{i,j}(t) = \sigma(t) + j\omega(t)$  is assigned, keep  $\omega(t)$  from vanishing.
4. The PD-eigenvalues should be continuously differentiable  $n - 1$  times.

### 6. Output/State Decoupling by PD-Eigenstructure Assignment

In this section, a design procedure for output / state decoupling using PD-eigenstructure assignment is developed based on Theorem 4.2. According to that theorem, to decouple the  $i$ th output  $y_i(t)$  from the  $j$ th closed-loop mode  $\rho_j(t)$ , one needs only to assign the  $j$ th column PD-eigenvector  $p_j(t)$  in such a way that  $p_j(t)$  is orthogonal to  $C_i^T(t)$  for all  $t$ . A natural question to ask is that to what extent the orientation  $p_j(t)$  can be adjusted. The following theorem addresses this issue.

**Theorem 6.1 (Parametrization of PD-eigenstructure)**

Let  $\{\rho_k(t)\}_{k=1}^n$  be a PD-spectrum for  $A(t)$  with an associated column PD-modal matrix  $P_0(t)$ . Then  $P(t) = P_0(t)S(t)$  is also a column PD-modal matrix for  $A(t)$  if and only if  $S(t)$  satisfies

$$\dot{S}(t) = \Upsilon(t)S(t) - S(t)\Upsilon(t) \quad (6.1)$$

with  $\det S(t) \neq 0$ , where  $\Upsilon(t) = \text{diag}[\rho_1(t), \rho_2(t), \dots, \rho_n(t)]$ . Moreover, the general solution of  $S(t)$  is given by

$$S(t) = e^{\int_0^t \Upsilon(\tau) d\tau} H e^{-\int_0^t \Upsilon(\tau) d\tau} \quad (6.2)$$

where  $H \in \mathbb{C}^{n \times n}$  is an arbitrary nonsingular constant matrix, and

$$e^{\int_0^t \Upsilon(\tau) d\tau} = \text{diag}\left[e^{\int_0^t \rho_1(\tau) d\tau}, e^{\int_0^t \rho_2(\tau) d\tau}, \dots, e^{\int_0^t \rho_n(\tau) d\tau}\right] \quad (6.3)$$

**Proof.** Suppose  $P(t) = P_0(t)S(t)$  is a PD-modal matrix for  $A(t)$ . Then we have

$$\dot{P}_0(t) = A(t)P_0(t) - P_0(t)\Upsilon(t) \quad (6.4)$$

and

$$[P_0(t)S(t)]^{(1)} = A(t)[P_0(t)S(t)] - [P_0(t)S(t)]\Upsilon(t) \quad (6.5)$$

But

$$\begin{aligned} [P_0(t)S(t)]^{(1)} &= \dot{P}_0(t)S(t) + P_0(t)\dot{S}(t) \\ &= [A(t)P_0(t) - P_0(t)\Upsilon(t)]S(t) + P_0(t)\dot{S}(t) \end{aligned} \quad (6.6)$$

Since  $\det P_0(t) \neq 0$ , it follows from (6.5) and (6.6) that

$$\dot{S}(t) = \Upsilon(t)S(t) - S(t)\Upsilon(t) .$$

Conversely, suppose that (6.1) holds. Then

$$\begin{aligned} [P_0(t)S(t)]^{(1)} &= \dot{P}_0(t)S(t) + P_0(t)\dot{S}(t) \\ &= [A(t)P_0(t) - P_0(t)T(t)]S(t) + P_0(t)[T(t)S(t) - S(t)T(t)] \\ &= A(t)[P_0(t)S(t)] - [P_0(t)S(t)]T(t) \end{aligned}$$

Thus  $P(t) = P_0(t)S(t)$  is a column PD-modal matrix. The general solution for  $S(t)$  given in (6.2) can be verified by direct computation.  $\square$

### Remark.

Theorem 6.1 is instrumental in the output/state decoupling by PD-eigenstructure assignment, where the orientation of some column PD-eigenvectors should be adjusted so as to keep orthogonal to a row vector in the output measurement matrix  $C(t)$ . Theorem 6.1 points out that the orientation of a column eigenvector is determined by: (i) the constant matrix  $H$ , and (ii) the associated PD-spectrum  $\{\rho_k(t)\}_{k=1}^n$ . Note that  $H = P_0^{-1}(0)P(0)$  is fixed at  $t = 0$ . Thus an optimal  $H$  should be computed off-line to achieve the desired PD-eigenvector orientations for all permissible PD-spectra. Then an on-line, real-time optimization is needed to select the optimal PD-spectrum to achieve the best tracking performance and decoupling. This design approach is illustrated in Figures 6.1 and 6.2.

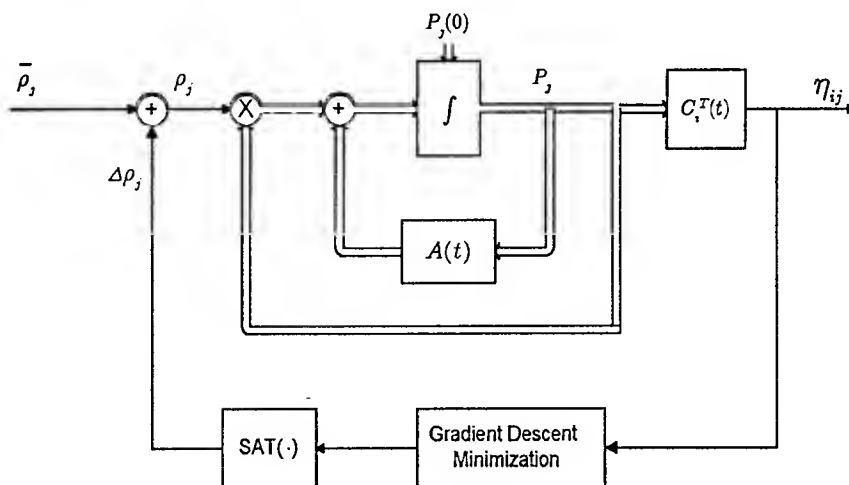


Figure 6.1 Output Decoupling by PD-Eigenstructure Assignment

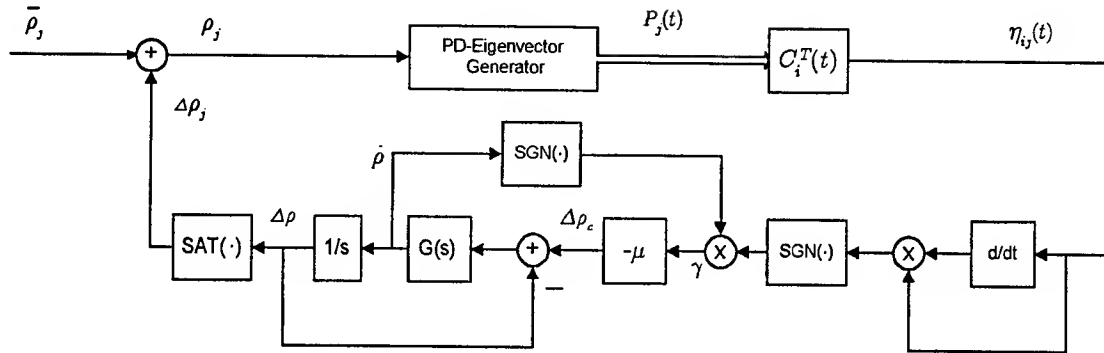


Figure 6.2 PD-Eigenstructure Assignment Logic for Output Decoupling

#### PD-Eigenstructure Assignment Procedure:

1. Follow the PD-Eigenspectrum Assignment Procedure of Section 5 to achieve stability.
2. To associate the  $i$ th output with, and only with the  $j$ th mode of the closed-loop system, assign the  $j$ th PD-eigenvalue  $\rho_j(t)$  such that the  $i$ th row vector  $\mathbf{c}_i^T(t)$  of  $\mathbf{C}(t)$  becomes an associated row PD-eigenvector, *i.e.*

$$\mathcal{Q}_{[A+BK(\rho_j)-\rho_j I]}\mathbf{c}_i = 0 \quad (3)$$

where  $\mathbf{K}(\rho_j(t), t) = \mathbf{K}_p(\rho_j(t), t)\mathbf{L}^{-1}(t)$ . In particular, if  $\mathbf{c}_i(t) \equiv \mathbf{c}_i$  is constant, than (3) can be written as

$$\frac{\mathbf{c}_i^T[A(t) + \mathbf{B}(t)\mathbf{K}(\rho_j(t), t)]\mathbf{c}_i}{\|\mathbf{c}_i\|^2} = \rho_j(t) \quad (4)$$

3. To decouple the  $i$ th output from the  $j$ th mode of the closed-loop system, assign  $\rho_j(t)$  such that the  $i$ th row vector  $\mathbf{c}_i^T(t)$  of  $\mathbf{C}(t)$  is orthogonal to the  $j$ th column PD-eigenvector  $\mathbf{p}_j(t)$ , *i.e.*

$$\mathcal{P}_{[A+BK(\rho_j)-\rho_j I]}\mathbf{p}_j = 0 \quad (5)$$

and

$$\mathbf{c}_i^T(t)\mathbf{p}_j(t) \equiv 0, \quad \mathbf{p}_j(t) \neq 0 \quad (6)$$

Remarks.

1. For eigenstructure assignment in a LTI system, once the feedback control gain  $\mathbf{K}$  is determined, the decoupling condition (6) is fixed, because the column eigenvector  $\mathbf{p}_j$  is fixed up to a constant scaling factor which does not change the orientation of  $\mathbf{p}_j$ . Whereas for PD-eigenstructure assignment, condition (6) can be optimized by selecting an optimal initial condition  $\mathbf{p}_j(0)$  for (5) and a permissible  $\rho_j(t)$ .
2. For eigenstructure assignment in a LTI system with  $l$  control inputs, only  $n$  of the  $n \times l$  gains in the feedback gain matrix  $\mathbf{K}$  are need for eigenvalue assignment, the rest  $n \times (l - 1)$  gains can be used to alter the orientations of the eigenvectors for optimal decoupling. However, at the present time, the eigenvalue assignment based on Theorem 5.2 requires the use of all  $n \times l$  gains, leaving no freedom in eigenvector assignment. However, this is not an intrinsic limitation. Additional research is needed to alleviate this limitation.

## 7. Roll-Yaw Decoupled Autopilot for a BTT Missile

In this section, the results obtained in Sections 2-6 are applied to the autopilot design for command tracking and roll-yaw decoupling of the EMRAAT bank-to-turn missile airframe. The complete nonlinear state equation of the EMRAAT airframe is given in the Appendix. For simplicity, it is assumed here that the roll-yaw dynamics are sufficiently decoupled from the pitch dynamics.

The basic design approach is by linearization of the nonlinear roll-yaw airframe along a nominal trajectory as follows.

Let the system states and inputs be chosen as

$$\xi(t) = \begin{bmatrix} \xi_1(t) \\ \xi_2(t) \\ \xi_3(t) \\ \xi_4(t) \end{bmatrix} = \begin{bmatrix} \beta(t) \\ r(t) \\ p(t) \\ \Phi(t) \end{bmatrix}$$

$$\delta(t) = \begin{bmatrix} \delta_r(t) \\ \delta_p(t) \end{bmatrix}$$

where  $\beta(t)$ ,  $r(t)$ ,  $p(t)$ ,  $\Phi(t)$  are the sideslip, yaw rate, roll rate and roll angle, respectively,  $\delta_r(t)$  and  $\delta_p(t)$  are the rudder and aileron angles, respectively. Since all state variables are on-line measurable, they are defined as the outputs. Thus, the measurement matrix  $C(t) \equiv I$ . The design objectives are to decouple  $\beta(t)$  from  $p(t)$  while maintaining good command tracking performance.

Then the non-linear state equation is given by

$$\dot{\xi} = f(\xi, \delta) = \begin{bmatrix} f_1(\xi_1, \xi_2, \xi_3, \xi_4, \delta_r, \delta_p) \\ f_2(\xi_1, \xi_2, \xi_3, \xi_4, \delta_r, \delta_p) \\ f_3(\xi_1, \xi_2, \xi_3, \xi_4, \delta_r, \delta_p) \\ f_4(\xi_1, \xi_2, \xi_3, \xi_4, \delta_r, \delta_p) \end{bmatrix}$$

For a nominal trajectory  $\bar{\xi}$  satisfying

$$\dot{\bar{\xi}}(t) = f(\bar{\xi}(t), \bar{\delta}(t))$$

define the tracking errors by

$$x(t) = \xi(t) - \bar{\xi}(t)$$

and the tracking error control input by

$$u(t) = \delta(t) - \bar{\delta}(t)$$

Then the linearized tracking error dynamics are given by

$$\dot{x} = A(t)x + B(t)u$$

$$y = C(t)x + D(t)u$$

For the EMRAAT airframe at altitude of 30,000 ft. and Mach of 2.0 [17]

$$A(t) = \frac{\partial f}{\partial \xi} \Big|_{\bar{\xi}, \bar{\delta}} = \begin{bmatrix} a_{11}(t) & a_{12}(t) & a_{13}(t) & a_{14}(t) \\ 96.09 & a_{22}(t) & a_{23}(t) & 0 \\ 1001 & a_{32}(t) & a_{33}(t) & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

$$B(t) = \frac{\partial f}{\partial \delta} \Big|_{\bar{\xi}, \bar{u}} = \begin{bmatrix} 1.979 \cdot 10^{-3} \cos(\bar{\beta}) & -247.3 \cdot 10^{-6} \cos(\bar{\beta}) \\ -75.99 & 17.52 \\ -959.5 & -1244 \\ 0 & 0 \end{bmatrix}$$

$$C(t) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$D(t) = 0$$

where

$$a_{11}(t) = \frac{gQS}{WV} \{ C_{Y_\beta} \cos(\bar{\beta}(t)) - (C_{Y_\beta} \bar{\beta}(t) + C_{Y_p} \bar{p}(t) + C_{Y_r} \bar{r}(t) + \\ + C_{Y_{\delta_p}} \bar{\delta}_p(t) + C_{Y_{\delta_r}} \bar{\delta}_r(t)) \sin(\bar{\beta}(t)) \} - \frac{g}{V} \sin(\bar{\phi}(t)) \sin(\bar{\beta}(t))$$

$$a_{12}(t) = -1 + 398.21 \cdot 10^{-6} \cos(\bar{\beta}(t))$$

$$a_{13}(t) = -18.056 \cdot 10^{-6} \cos(\bar{\beta}(t))$$

$$a_{14}(t) = 16.631 \cdot 10^{-3} \cos(\bar{\phi}(t)) \cos(\bar{\beta}(t))$$

$$a_{22}(t) = 0.3623 \cdot 10^{-3} \bar{r}(t) + 0.5024 \cdot 10^{-3} \bar{p}(t) - 0.6053$$

$$a_{23}(t) = 7.8 \cdot 10^{-3} \bar{p}(t) + 0.5024 \cdot 10^{-3} \bar{r}(t) - 3.5 \cdot 10^{-3}$$

$$a_{32}(t) = -36.8 \cdot 10^{-3} \bar{r}(t) - 2.3 \cdot 10^{-3} \bar{p}(t) + 0.8056$$

$$a_{33}(t) = 35.80 \cdot 10^{-6} \bar{p}(t) - 2.3 \cdot 10^{-3} \bar{r}(t) - 2.177$$

The overall system configuration is shown in Figure 7.1. The PD-eigenstructure assignment control for LTV error dynamics stabilization and decoupling is shown in Figure 7.2.



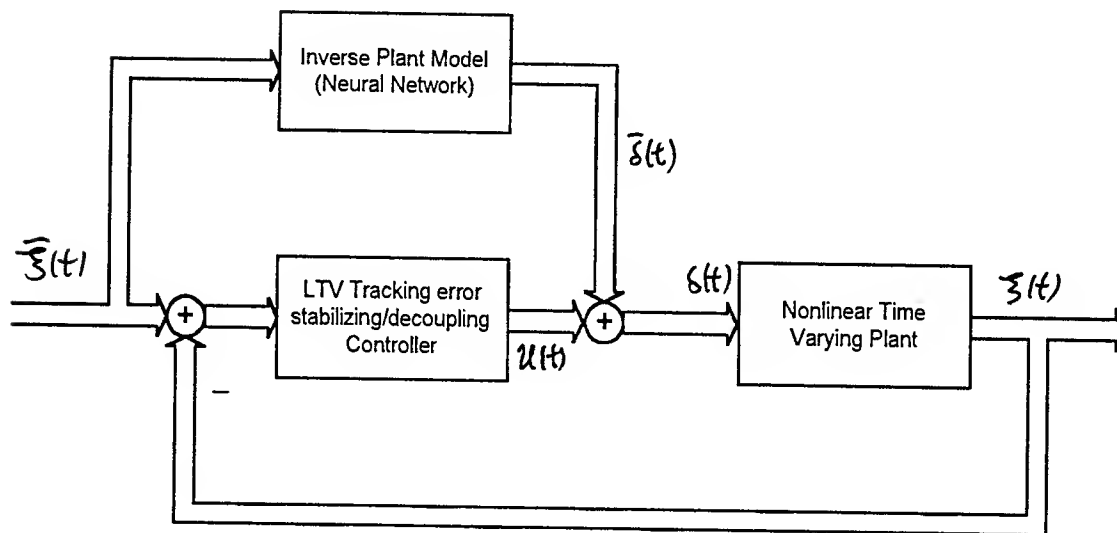


Figure 7.1 Nonlinear Tracking System Configuration

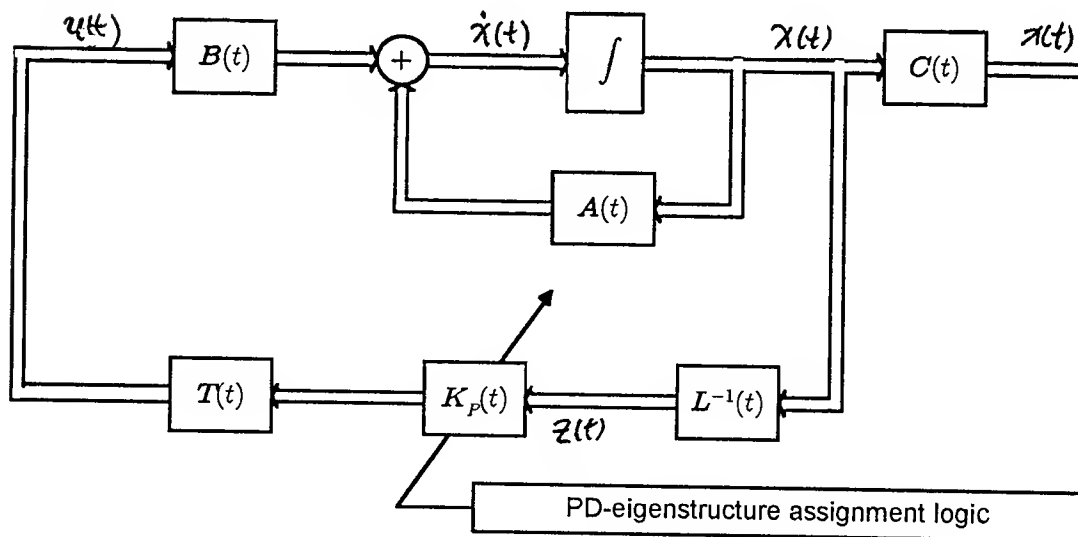


Figure 7.2 LTV Tracking Error Stabilizing and Decoupling Control

Now apply the S-W transformation to obtain the MVPV canonical form with lexicographic indices  $n_1 = 2, n_2 = 2$ :

$$A_p(t) = \left[ \begin{array}{cc|cc} 0 & 1 & 0 & 0 \\ \alpha_{21}(t) & \alpha_{22}(t) & \alpha_{23}(t) & \alpha_{24}(t) \\ 0 & 0 & 0 & 1 \\ \alpha_{41}(t) & \alpha_{42}(t) & \alpha_{43}(t) & \alpha_{44}(t) \end{array} \right]$$

$$B_p(t) = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}$$

Then design the state feedback control gain

$$K_p(t) = \left[ \begin{array}{cc|cc} \beta_{21}(t) - \alpha_{21}(t) & \beta_{22}(t) - \alpha_{22}(t) & -\alpha_{23}(t) & -\alpha_{24}(t) \\ -\alpha_{41}(t) & -\alpha_{42}(t) & \beta_{43}(t) - \alpha_{43}(t) & \beta_{44}(t) - \alpha_{44}(t) \end{array} \right]$$

to obtain the desired closed-loop dynamics in the MVPV coordinates

$$A_p(t) + B_p(t)K_p(t) = \left[ \begin{array}{cc|cc} 0 & 1 & 0 & 0 \\ \beta_{21}(t) & \beta_{22}(t) & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & \beta_{43}(t) & \beta_{44}(t) \end{array} \right]$$

where  $\beta_{ij}(t)$  are synthesized from the (real) PD-spectral canonical form

$$A_s(t) = \left[ \begin{array}{cc|cc} \sigma_1(t) & \omega_1(t) & 0 & 0 \\ -\omega_1(t) & \sigma_1(t) & 0 & 0 \\ 0 & 0 & \sigma_{21}(t) & 0 \\ 0 & 0 & 0 & \sigma_{22}(t) \end{array} \right]$$

with the desired closed-loop PD-eigenvalues  $\rho_{11,12}(t) = \sigma_1(t) \pm j\omega_1(t)$  and  $\rho_{21}(t) = \sigma_{21}(t)$ ,  $\rho_{22}(t) = \sigma_{22}(t)$ , where the roll modes are chosen to be simple (nonoscillatory).

Due to the time constraints on this research, implementation and simulation results are not available. Anticipated difficulties of the implementation are (i) the complexity of the control law  $k_p(t)$ , (ii) training of the neural network inverse of the unstable nonlinear airframe, and (iii) real-

time implementation of the eigenstructure assignment logic. However, none of these problems is insurmountable with further studies.

## **8. Summary and Conclusions**

In this research a systematic design procedure for nonlinear tracking and output/state decoupling control has been developed by way of linearization along a nominal trajectory. The resulting LTV tracking error dynamics are then stabilized and decoupled using PD-eigenstructure assignment in a way similar to the eigenstructure assignment design for LTI systems. Theoretical results on the assignability of the PD-eigenstructure for stabilization and decoupling have been obtained. However, due to the time constraint, implementation and simulation results are not available at the present. Further research is needed to address the complexity of the implementation, and to validation the theory and design procedure by simulations.

It is believed that this research is the first attempt at stabilization and decoupling of time-varying dynamics using a differential-algebraic approach, without resorting to the unreliable frozen-state, frozen-time gain scheduling. Preliminary results have shown excellent performance and robustness with a simple pitch autopilot. Thus, further study on the implementation of the significantly more complex is warranted. Furthermore, the PD-eigenstructure assignment control is by no means limited to stabilization and decoupling. An array of challenges posed by modern missile technology [2] can be addressed by the multiobjective PD-eigenstructure assignment concept illustrated in Figure 8.1. Additional applications can be found in aircraft flight control, spacecraft altitude control, vibration control, robotics and other nonlinear tracking and fast time-varying control systems.

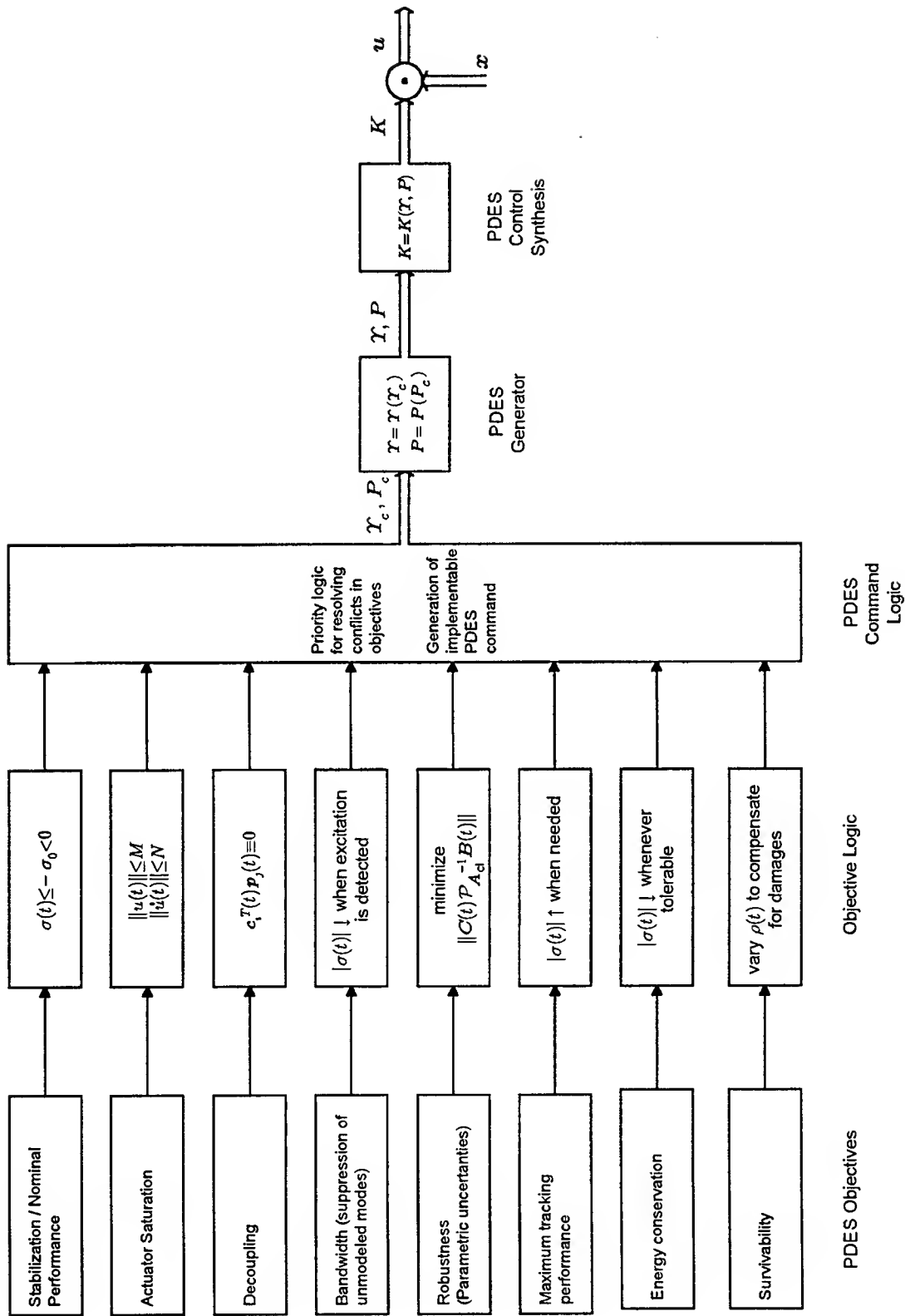


Figure 8.1 Multiobjective PD-Eigenstructure (PDES) Assignment Control

## Appendix

### State Equations for the EMRAAT Missile [17]

$$\begin{aligned}\dot{\alpha} = & q - \tan(\beta)[p(\cos(\alpha) - r\sin(\alpha))] + \frac{g}{V\cos(\beta)}(\cos(\alpha)\cos(\phi)\cos(\theta) + \sin(\alpha)\sin(\theta)) \\ & + \frac{gQS}{WV\cos(\beta)}(C_{N_\alpha}\alpha + C_{N_\alpha}\dot{\alpha} + C_{N_q}q + C_{N_\delta}\delta_q)\cos(\alpha)\end{aligned}$$

$$\begin{aligned}\dot{\beta} = & p\sin(\alpha) - r\cos(\alpha) + \frac{gQS}{WV}(C_{Y_\beta}\beta + C_{Y_p}p + C_{Y_r}r + C_{Y_{\delta_p}}\delta_p + C_{Y_{\delta_r}}\delta_r)\cos(\beta) \\ & + \frac{g}{V}\cos(\theta)\sin(\phi)\cos(\beta)\end{aligned}$$

$$\begin{aligned}\dot{p} = & [(-I_{xy}I_{xz}I_{zz} - I_{xz}^2I_{yz} + I_{xy}^2I_{yz} + I_{xy}I_{xz}I_{yy})p^2 + \\ & + (I_{yy}I_{yz}I_{zz} - I_{yz}^3 - I_{xy}^2I_{yz} - I_{xy}I_{xz}I_{yy})q^2 + \\ & + (-I_{yy}I_{yz}I_{zz} + I_{xy}I_{xz}I_{zz} + I_{yz}^3 + I_{xz}^2I_{yz})r^2 + \\ & + (-I_{xy}I_{yz}I_{zz} + I_{xz}I_{yy}I_{zz} - 2I_{yz}^2I_{xz} - I_{xy}I_{yy}I_{yz} + I_{xx}I_{xy}I_{yz} - I_{xz}I_{yy}^2 + I_{xx}I_{xz}I_{yy})pq + \\ & + (I_{xy}I_{zz}^2 + I_{xz}I_{yz}I_{zz} - I_{xy}I_{yy}I_{zz} - I_{xx}I_{xy}I_{zz} + 2I_{xy}I_{yz}^2 + I_{xz}I_{yy}I_{yz} - I_{xx}I_{xz}I_{yz})pr + \\ & + (-I_{yy}I_{zz}^2 + I_{yz}^2I_{zz} + I_{yy}^2I_{zz} - I_{yy}I_{yz}^2 - I_{xz}^2I_{yy})qr + \\ & + QSd(C_{l_p}(I_{yy}I_{zz} - I_{yz}^2) + C_{n_p}(I_{xy}I_{yz} + I_{xz}I_{yy}))p + \\ & + QSd(C_{m_q}(I_{xy}I_{zz} + I_{xz}I_{yz}))q + \\ & + QSd(C_{l_r}(I_{yy}I_{zz} - I_{yz}^2) + C_{n_r}(I_{xy}I_{yz} + I_{xz}I_{yy}))r + \\ & + QSd(C_{m_\alpha}(I_{xy}I_{zz} + I_{xz}I_{yz}))\dot{\alpha} + \\ & + QSd(C_{m_\alpha}(I_{xy}I_{zz} + I_{xz}I_{yz}))\alpha + \\ & + QSd(C_{l_\beta}(I_{yy}I_{zz} - I_{yz}^2) + C_{n_\beta}(I_{xy}I_{yz} + I_{xz}I_{yy}))\beta + \\ & + QSd(C_{l_{\delta_p}}(I_{yy}I_{zz} - I_{yz}^2) + C_{n_{\delta_p}}(I_{xy}I_{yz} + I_{xz}I_{yy}))\delta_p + \\ & + QSd(C_{m_{\delta_q}}(I_{xy}I_{zz} + I_{xz}I_{yz}))\delta_q + \\ & + QSd(C_{l_{\delta_r}}(I_{yy}I_{zz} - I_{yz}^2) + C_{n_{\delta_r}}(I_{xy}I_{yz} + I_{xz}I_{yy}))\delta_r] \\ & + (I_{xx}I_{yy}I_{zz} - I_{xy}^2I_{zz} - I_{xx}I_{yz}^2 - 2I_{xy}I_{xz}I_{yz} - I_{xz}^2I_{yy})^{-1}\end{aligned}$$

$$\begin{aligned}
\dot{q} = & [(-I_{xx}I_{xz}I_{zz} + I_{xx}I_{xy}I_{yz} + I_{xy}^3 + I_{xy}^2I_{xz})p^2 + \\
& + (I_{xy}I_{yz}I_{zz} + I_{yz}^2I_{xz} - I_{xx}I_{xy}I_{yz} - I_{xy}^2I_{xz})q^2 + \\
& + (-I_{xy}I_{yz}I_{zz} + I_{xx}I_{xz}I_{zz} - I_{xz}I_{yz}^2 - I_{xz}^3)r^2 + \\
& + (-I_{xx}I_{yz}I_{zz} + I_{xy}I_{xz}I_{zz} + 2I_{xz}^2I_{yz} - I_{xx}I_{yy}I_{yz} - I_{xy}I_{xz}I_{yy} + I_{yz}I_{xx}^2 + I_{xx}I_{xy}I_{xz})pq + \\
& + (I_{xx}I_{zz}^2 - I_{xz}^2I_{zz} - I_{xy}^2I_{zz} - I_{xz}^2I_{zz} + I_{xx}I_{yz}^2 + I_{xz}^2I_{xx})pr + \\
& + (-I_{xy}I_{zz}^2 - I_{xz}I_{yz}I_{zz} + I_{xy}I_{yy}I_{zz} + I_{xx}I_{xy}I_{zz} + I_{xz}I_{yy}I_{yz} - I_{xz}I_{xx}I_{yz} - 2I_{xz}^2I_{xy})qr + \\
& + Q\dot{S}d(C_{l_p}(I_{xy}I_{zz} - I_{xz}I_{yz}) + C_{n_p}(I_{xx}I_{yz} + I_{xy}I_{xz}))p + \\
& + Q\dot{S}d(C_{m_q}(I_{xx}I_{zz} - I_{xz}^2))q + \\
& + Q\dot{S}d(C_{l_r}(I_{xy}I_{zz} + I_{xz}I_{yz}) + C_{n_r}(I_{xx}I_{yz} + I_{xy}I_{xz}))r + \\
& + Q\dot{S}d(C_{m_\alpha}(I_{xx}I_{zz} - I_{xz}^2))\dot{\alpha} + \\
& + Q\dot{S}d(C_{m_\alpha}(I_{xx}I_{zz} - I_{xz}^2))\alpha + \\
& + Q\dot{S}d(C_{l_\beta}(I_{xy}I_{zz} + I_{xz}I_{yz}) + C_{n_\beta}(I_{xx}I_{yz} + I_{xz}I_{xy}))\beta + \\
& + Q\dot{S}d(C_{l_\delta p}(I_{xy}I_{zz} + I_{xz}I_{yz}) + C_{n_\delta p}(I_{xx}I_{yz} + I_{xz}I_{xy}))\delta_p + \\
& + Q\dot{S}d(C_{m_{\delta q}}(I_{xx}I_{zz} - I_{xz}^2))\delta_q + \\
& + Q\dot{S}d(C_{l_\delta r}(I_{xy}I_{zz} + I_{xz}I_{yz}) + C_{n_\delta r}(I_{xx}I_{yz} + I_{xz}I_{xy}))\delta_r] \\
& (I_{xx}I_{yy}I_{zz} - I_{xy}^2I_{zz} - I_{xx}I_{yz}^2 - 2I_{xy}I_{xz}I_{yz} - I_{xz}^2I_{yy})^{-1}
\end{aligned}$$

$$\begin{aligned}
\dot{r} = & [(-I_{xx}I_{xz}I_{yz} + I_{xx}I_{xy}I_{yz} - I_{xy}^3 - I_{xz}^2I_{xy})p^2 + \\
& + (I_{xz}I_{yy}I_{yz} + I_{yz}^2I_{xy} - I_{xx}I_{xy}I_{yz} + I_{xy}^3)q^2 + \\
& + (-I_{xy}I_{yz}^2 - I_{xz}I_{yy}I_{yz} + I_{xx}I_{xz}I_{yz} + I_{xy}I_{xz}^2)r^2 + \\
& + (-I_{xx}I_{yz}^2 - I_{yz}^2I_{xx} + I_{xz}^2I_{yy} + I_{xy}^2I_{yz} + I_{yy}I_{xx}^2 - I_{xy}^2I_{xz})pq + \\
& + (I_{xx}I_{yz}I_{zz} + I_{xy}I_{xz}I_{zz} + I_{xx}I_{yy}I_{yz} - 2I_{xz}^2I_{yz} - I_{yz}I_{xx}^2 - I_{xy}I_{xz}I_{yy} - I_{xx}I_{xy}I_{xz})pr + \\
& + (-I_{xy}I_{yz}I_{zz} - I_{xz}I_{yy}I_{zz} + I_{xy}I_{yy}I_{yz} + I_{xx}I_{xy}I_{yz} + I_{xz}I_{yy}^2 - I_{xx}I_{xz}I_{yy} + 2I_{xz}^2I_{xy})qr + \\
& + Q\dot{S}d(C_{l_p}(I_{xy}I_{yz} - I_{xz}I_{yy}) + C_{n_p}(I_{xx}I_{yz} - I_{xy}^2))p + \\
& + Q\dot{S}d(C_{m_q}(I_{xx}I_{yz} + I_{xy}I_{xz}))q + \\
& + Q\dot{S}d(C_{l_r}(I_{xy}I_{yz} + I_{xz}I_{yy}) + C_{n_r}(I_{xx}I_{yz} - I_{xy}^2))r + \\
& + Q\dot{S}d(C_{m_\alpha}(I_{xx}I_{yz} - I_{xy}I_{xz}))\dot{\alpha} + \\
& + Q\dot{S}d(C_{m_\alpha}(I_{xx}I_{yz} - I_{xy}I_{xz}))\alpha + \\
& + Q\dot{S}d(C_{l_\beta}(I_{xy}I_{yz} + I_{xz}I_{yy}) + C_{n_\beta}(I_{xx}I_{yz} - I_{xy}^2))\beta + \\
& + Q\dot{S}d(C_{l_\delta p}(I_{xy}I_{yz} + I_{xz}I_{yy}) + C_{n_\delta p}(I_{xx}I_{yz} - I_{xy}^2))\delta_p + \\
& + Q\dot{S}d(C_{m_{\delta q}}(I_{xx}I_{yz} - I_{xy}I_{xz}))\delta_q + \\
& + Q\dot{S}d(C_{l_\delta r}(I_{xy}I_{yz} + I_{xz}I_{yy}) + C_{n_\delta r}(I_{xx}I_{yz} - I_{xy}^2))\delta_r] \\
& (I_{xx}I_{yy}I_{zz} - I_{xy}^2I_{zz} - I_{xx}I_{yz}^2 - 2I_{xy}I_{xz}I_{yz} - I_{xz}^2I_{yy})^{-1}
\end{aligned}$$

## Glossary of Terms

$\alpha$	— Angle of attack
$\beta$	— Angle of sideslip
$p$	— Roll rate
$q$	— Pitch rate
$r$	— Yaw rate
$Q$	— Dynamic Pressure
$S$	— Reference area
$d$	— Reference length/diameter
$V$	— Missile velocity
$W$	— Missile weight
$g$	— Acceleration due to gravity
$\psi$	— Yaw angle
$\theta$	— Pitch angle
$\phi$	— Roll angle
$\delta p$	— Roll control input (surface deflection)
$\delta q$	— Pitch control input (surface deflection)
$\delta r$	— Yaw control input (surface deflection)
$\tilde{Q}$	— $(gQS/WV)$
$N$	— Normal force
$Y$	— Side force
$C_{a_b}$	— Aerodynamic Coefficient- $a$ due to $b$
$l$	— Aerodynamic moment about $x$ -axis
$m$	— Aerodynamic moment about $y$ -axis
$n$	— Aerodynamic moment about $z$ -axis
$I_{ij}$	— Moment or product of inertia

## Aerodynamic coefficients (Mach = 2.0)

$C_{N_\alpha}$	= 36.6
$C_{n_\alpha}$	= 0.0274
$C_{N_q}$	= 0.0145
$C_{N_{\dot{\delta}_q}}$	= 6.0165
$C_{Y_\beta}$	= -14.9
$C_{Y_p}$	= -0.00073
$C_{Y_r}$	= 0.0161
$C_{Y_{\dot{\delta}_p}}$	= -.01
$C_{Y_{\dot{\delta}_r}}$	= .08
$C_{l_\beta}$	= 5.44
$C_{l_p}$	= -0.011
$C_{l_r}$	= 0.0021
$C_{l_{\dot{\delta}_p}}$	= -6.30
$C_{l_{\dot{\delta}_r}}$	= -5.16
$C_{m_\alpha}$	= -82
$C_{m_{\dot{\alpha}}}$	= -0.014
$C_{m_q}$	= -0.202
$C_{m_{\dot{\delta}_q}}$	= -40.7
$C_{n_\beta}$	= 35.52
$C_{n_p}$	= 0.006
$C_{n_r}$	= -0.2
$C_{n_{\dot{\delta}_p}}$	= 1.72
$C_{n_{\dot{\delta}_r}}$	= -28.65

## EMRAAT Physical Properties

$g$	= 32.2 ft/s <sup>2</sup>
$d$	= 0.625 ft.
$S$	= 0.3067 ft <sup>2</sup>
$W$	= 227 lbs
$V$	= 1936.16 ft/s
$M$	= 2.0
$Q$	= 1100.75 lb/ft <sup>2</sup>
$I_{xx}$	= 1.08 slug*ft <sup>2</sup>
$I_{yy}$	= 70.13 slug*ft <sup>2</sup>
$I_{zz}$	= 70.66 slug*ft <sup>2</sup>
$I_{xy}$	= 0.274 slug*ft <sup>2</sup>
$I_{xz}$	= -0.704 slug*ft <sup>2</sup>
$I_{yz}$	= 0.017 slug*ft <sup>2</sup>
Air Density	= $5.87 \times 10^{-4}$ slug/ft <sup>3</sup>
Altitude	= 30,000 ft.

## References

- [1] C. N. D'Souza and C. P. Mracek, "Derivation of the Full Nonlinear Equations of Motion for a Rigid Airframe," *To appear in AIAA J. of Guidance, Control and Dynamics*, 1995.
- [2] J. R. Cloutier, J. H. Evers and J. J. Feeley, "Assessment of Air-to-Air Missile Guidance and Control Technology," *IEEE Control Systems Magazine*, Oct. 1989.
- [3] R. F. Wilson, J. R. Cloutier and R. K. Yedavalli, "Control Design for Robust Eigenstructure Assignment in Linear Uncertain Systems," *IEEE Control Systems Magazine*, Oct. 1992.
- [4] J. S. Shamma and M. Athans, "Gain Scheduling: Potential Hazards and Possible Remedies," *IEEE Control System Magazine*, 101-107, June 1992.
- [5] L. H. Carter, "Gain-Scheduled Bank-to-Turn Missile Auto-Pilot Design Using Linear Parameter Varying Transformations," *Final Reports for the 1994 AFOSR Graduate Student Research Program*, pp. 5-1 to 5-20, Sept., 1994.
- [6] J. Zhu, "A unified spectral theory for linear time-varying systems—progress and challenges," *Proceedings, 34th IEEE Conference on Decision and Control*, New Orleans, LA, 2540-2546, Dec. 1995.
- [7] J. Zhu, "Missile Autopilot Design Based on a Unified Spectral Theory for Linear Time-Varying Systems," *Final Reports for the 1995 AFOSR SFRP*, pp. 64-1 to 64-20, July, 1995.
- [8] J. Zhu and M. C. Mickle, "Missile autopilot design using a new linear time-varying control technique," *AIAA Journal on Guidance, Control and Dynamics*, vol. 20, no. 1, 1-8, Jan-Feb. 1997.
- [9] M. G. Floquet, "Sur la Théorie des Équations Différentielles Linéaires," *Annales Scientifiques de L'École Normale Supérieure*, 1879.
- [10] E. L. Ince, "Ordinary Differential Equations", Dover Publications, Inc., New York, 1956 (1st ed. 1926).
- [11] J. Zhu, "A Unified Eigenvalue Theory for Linear Dynamical Systems", Ph. D. Dissertation, Electrical and Computer Engineering Department, University of Alabama in Huntsville, May 1989.
- [12] L. M. Silverman and B. D. O. Anderson, "Controllability, Observability and Stability of Linear Systems," *Siam Journal of Control*, vol. 6, no. 1, pp. 121-130, 1968.
- [13] W. A. Wolovich, "On the Stabilization of Controllable Systems," *IEEE Transactions on Automatic Control*, vol. 13, pp. 569-572, October 1968.
- [14] T. Kailath, "Linear Systems", Prentice-Hall, 1980.



- [15] C. E. Seal and A. R. Stubberud, "Canonical Forms for Multi-Input Time-Variable Systems," *IEEE Transactions on Automatic Control*, vol. 14, pp. 704-707, December 1969.
- [16] J. Zhu, "A necessary and sufficient stability criterion for linear time-varying systems," *Proceedings, 28th IEEE Southeastern Symposium on Systems Theory*, 115-119, Baton Rouge, LA, April 1996..
- [17] D. A. Schumacher, "Tactical Missile Autopilot Design Using Nonlinear Control", Ph. D. Dissertation, Aerospace Engineering Department, University of Michigan, 1994.

**Acknowledgment:**

The author gratefully acknowledge the Air Force Office of Scientific Research, Bolling Air Force Base, and the Wright Laboratory, Eglin Air Force Base for financial support of this research. The author sincerely thanks Dr. J. Cloutier, who served as the focal point of this research, Colonel C. Mracek, Dr. R. Haul of the MNAG branch for their inspiration and valuable discussions during this work. The author also appreciates technical assistance from Mr. C. M. Mickle and Ms. Y. M. Zhu of the Automatix, Inc., Baton Rouge, Louisiana.

Design and Implementation of a Global Navigation Satellite System (GNSS)  
Software Radio Receiver

Dennis M. Akos  
Graduate Research Assistant  
Department of Electrical Engineering

Ohio University  
Athens, OH 45701

Final Report for:  
Summer Research Extension Program  
Wright Laboratory

Sponsored by:  
Air Force Office of Scientific Research  
Bolling AFB, Washington, DC

and

Wright Laboratory

December 1996

---

DESIGN AND IMPLEMENTATION OF A GLOBAL NAVIGATION SATELLITE SYSTEM (GNSS)  
SOFTWARE RADIO RECEIVER

Dennis M. Akos  
Ph.D. Candidate  
Department of Electrical Engineering and Computer Science  
Ohio University

Abstract

A prototype Global Navigation Satellite System (GNSS) software radio has been successfully developed. A software radio has many advantages over the architecture of a traditional receiver. These include a tighter integration between simulation and implementation, a tremendous level of versatility in the final design, and the ability for a single receiver to function as multiple receivers. The focus of this implementation, a GNSS receiver, is a navigation receiver and will bring all the benefits of the software radio to the navigation community. The preliminary work accomplished in the development of the GNSS software radio thus far is the implementation of the receiver front end, data collection hardware, and signal processing algorithms. This work has resulted in a postprocessed position solution within 500 meter solely through the use of software-based signal processing.

# DEVELOPMENT OF A GLOBAL NAVIGATION SATELLITE SYSTEM SOFTWARE RADIO

Dennis M. Akos

## Introduction

The software radio describes a receiver in which the majority of the signal processing is accomplished via a programmable microprocessor as opposed to analog or hardwired discrete components. This allows for a tighter integration of simulation and implementation as well as tremendous flexibility in the final design.

The software radio concept is being applied in the design of a GNSS receiver. However, this concept is not limited to the GNSS signal and could be expanded to include other navigation signals in the same radio design. This initial work will bring the benefits of such an implementation to the navigation community.

The paper begins by describing the ideal software radio and details its multiple benefits. The target implementation and the development testbed is characterized along with the necessary design steps. Finally, an informal discussion of various GNSS acquisition and tracking methodologies implemented is presented. Described here are those techniques implemented for use with the software radio and validated using actual GNSS data.

## Software Radio

There are two primary design goals in developing a software radio. First, the analog-to-digital converter (ADC) should be positioned as close to the antenna as possible in the front end of the receiver. Second, the resulting samples should be processed using a programmable microprocessor. These two principles provide all the benefits associated with the software radio.

Moving the ADC closer to the antenna in the RF front end chain eliminates additional components used in frequency translation. These components include: local oscillators (LO), mixers, and filters, all of which can contribute potential nonlinear effects as well as temperature and age based performance variations. Ideally, the receiver front end would consist of the antenna, amplifier, bandpass filter, and ADC. Frequency translation, since it is impractical to process the signal at RF, is accomplished via bandpass sampling [1].

Bandpass sampling is the process of sampling an information signal based on its bandwidth as opposed to its RF carrier. This has been proposed and implemented successfully with the GPS-SPS transmission [2]. Reference 2 details a front end design consisting of an antenna, amplifiers, filters and an ADC that sampled the 1575.42 MHz RF carrier directly at a rate of 5 MHz and achieved the desired frequency translation via bandpass sampling.

Processing the resulting ADC samples strictly in software provides additional benefits for the software radio concept. First, since all signal processing is accomplished in software there is tighter integration between simulation and actual receiver operation. If the signal degradations can be adequately simulated, performance can be accurately predicted using the actual discrete signal processing that will occur in the receiver. This is especially true now that the front end contains fewer possible error sources. Second, there is a tremendous level of flexibility in the receiver

design since all signal processing is software based. In order to incorporate the latest theoretical developments, costly hardware prototypes no longer need to be fabricated, rather they can be incorporated into the programming and evaluated. Various receiver architectures can be assessed simply by downloading the appropriate software to the target processor.

The front end and the flexibility of achieving all signal processing in software allow the design to serve as multiple radios [3]. Currently, there exist receivers, where the LO is adjusted to downconvert and process different frequency transmissions. These designs, however, are limited in their signal processing as a result of their hardwired architectures. With the software radio design a continuous range of frequencies could be captured by using a high sampling rate. Specific transmissions could then be digitally filtered out and processed. By changing the software processing, a single configuration could serve as an FM, AM, or PM receiver. This concept would be extremely beneficial in the navigation community as a single receiver could process and integrate multiple navigation signals for improved accuracy, reliability, and integrity.

Some of these ideas are reflected in the current software radio research. Reference 4 discusses a GPS L1/L2 front end design which bandpass samples the frequency band 1.2 - 1.6 GHz to utilize both GPS frequency transmissions. This front end is followed by digital filters used to extract the exact information bands of interest for further processing.

The software radio concept is not without its disadvantages, unfortunately. There are two primary technological factors which limit its current practicality. They are the current state-of-the-art in ADC technology and programmable processing power. For the bulk of the navigation community, the highest RF signals of interest are for the GNSS band which are below 1.7 GHz. ADC's do exist which can provide multi-bit sampling at rates up to 4 Gsps [5]. At this sampling rate all frequency information from DC to 2 GHz can be captured. However, programmable processing power significantly reduces the maximum possible data rate from that allowed by the ADC [6]. This leads to the capture of partial frequency bands, as in the case of bandpass sampling. The limits imposed by the current generation of programmable processors are so restrictive, a variation of bandpass sampling has been proposed for the combined digitization and processing of GPS-SPS and GLONASS signals [7].

These disadvantages are only temporary. ADC performance already exceeds what is required for the navigation community. The lag in available processor power should be eliminated in the near future. Moore's law, which has held true since the inception of the microprocessor, has shown processing power to double every 18 months.

#### GNSS SOFTWARE RADIO DEVELOPMENT

A navigation software radio will provide significant advantages over existing receivers and their traditional design. Although the current technology will not allow the development of an all-encompassing navigation software radio, the initial goal is the development of a GNSS software radio. As technology advances, the framework (front end hardware and software algorithms) will be in place to take advantage of increased processing power.

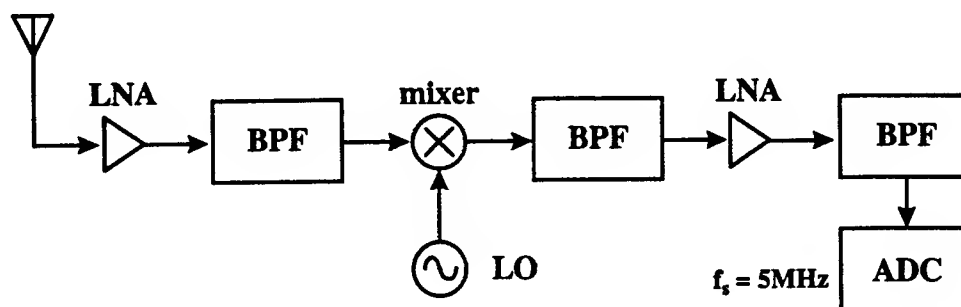


Figure 1. Front End Configuration used for Raw Data Collection

The development is planned for three stages. First, a front end utilizing multiband bandpass sampling was designed and implemented [7]. This research demonstrated proof of concept and a final design is under development. The second stage, currently underway, is the programming of the software algorithms necessary for processing of the sampled data. It is impractical to attempt to initially develop the algorithms to operate in real time. Rather a data set will be collected and postprocessed using the developing algorithms. The third stage will be the optimization of these algorithms to operate in real time. The target processor for the real time implementation is the Texas Instruments TMS320C80 DSP, one of the most powerful DPS processors available, capable of 2 BOPS.

In order to validate the spread spectrum acquisition and tracking algorithms it is necessary to obtain an adequate length data set. The difficulty lies in the pre-correlation bandwidth of the GNSS signal. GPS-SPS has a null-to-null bandwidth of approximately 2 MHz, therefore the minimum sampling frequency must be at least 4 MHz. A sampling frequency of 5 MHz is used to adequately capture the required frequency information. Assuming 8 bit samples, 30 seconds of data (a full navigation frame for GPS-SPS) requires 150 MB of storage space. In order to minimize the storage requirement, data sets of 12 seconds (or a two subframes) are collected for postprocessing. If subframes #1, 2, and 3 can be collected, the resulting data will contain enough information to establish a position solution, which is the primary purpose in a navigation receiver. This can be accomplished by collecting a single data set corresponding to the desired subframe and then storing that on a hard drive or an alternative long term storage device, then collecting the next desired subframe soon after. This process is repeated until all required subframes can be collected.

The data collection platform uses a more traditional front end design to reduce the requirement on the ADC. The configuration, depicted in Figure 1, employs a single downconversion stage to 21.25 MHz, where the signal is bandpass sampled at 5 MHz, resulting in a final IF of 1.25 MHz. This arrangement is utilized to collect GPS-SPS data sets and the development of generic GNSS signal processing algorithms. The data collection hardware is a Peripheral Component Interface (PCI) card for use with Intel-based microcomputers. The card utilizes two ADCs which allows 12-bit sampling at rates up to 60 MHz, more than adequate for the 2 MHz null-to-null bandwidth of the GPS-SPS signal. One distinct advantage of this card is the ability to use the PCI bus to write samples directly to the memory of the host PC as opposed to using expensive memory on the ADC card itself. In this configuration it is

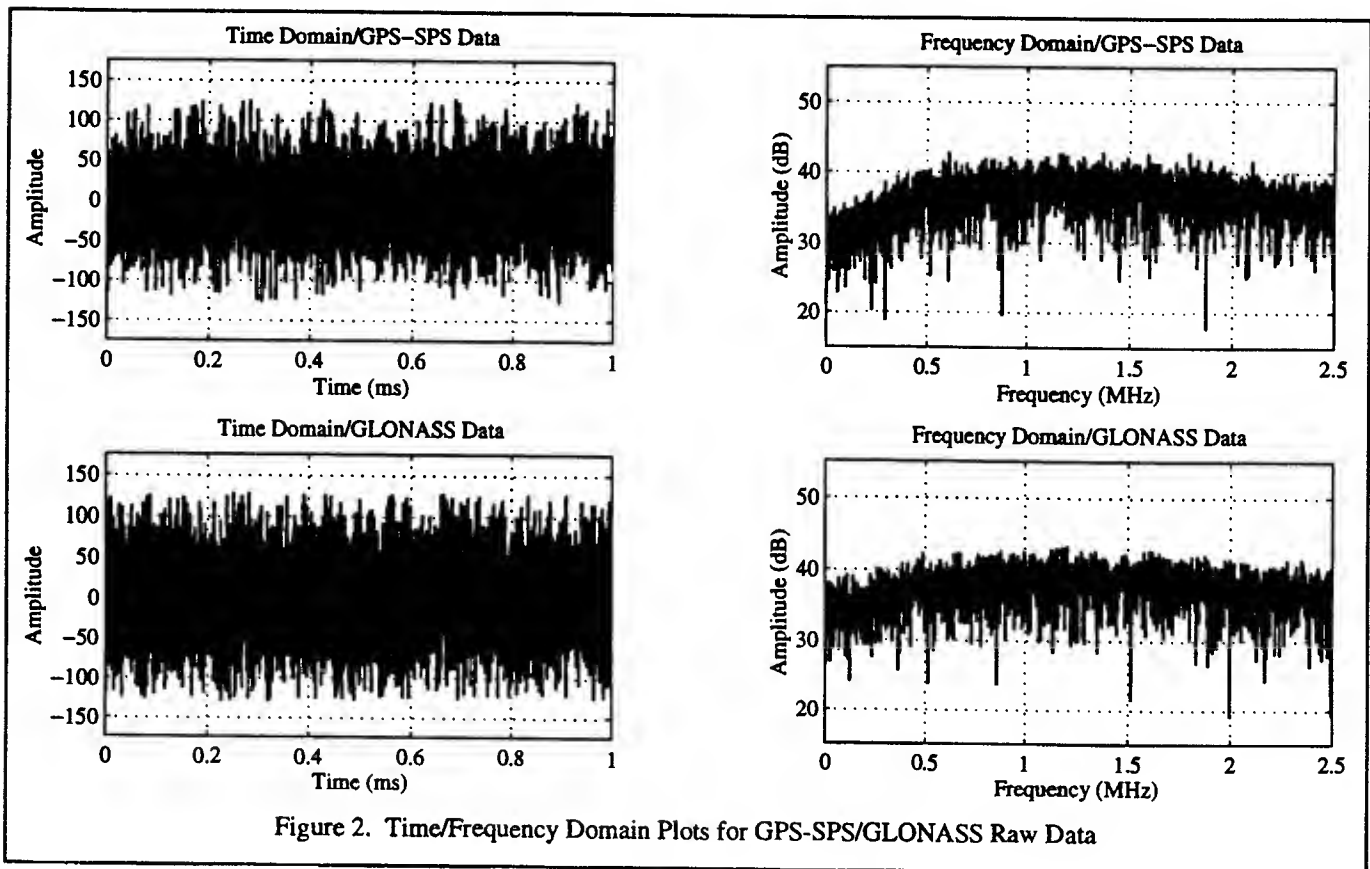


Figure 2. Time/Frequency Domain Plots for GPS-SPS/GLONASS Raw Data

possible to store a continuous data record of up to 128 MB, the maximum memory size in typical motherboards. Multiple data records of this length allow for capture of sufficient GNSS navigation data to solve for a position solution.

#### GNSS SOFTWARE RADIO ACQUISITION

The first stage in processing the code division multiple access (CDMA) format of the GPS-SPS navigation signal is acquisition. The spread spectrum modulation format essentially conceals any discernible signal in the raw data set when viewed in either the time or frequency domains. This is also true for GLONASS even though it employs frequency division multiple access (FDMA). Each frequency channel uses the same maximal length code as a spreading sequence for the purpose of time transfer. Raw data collected using the front end in Figure 1 is shown in Figure 2 in both the time and frequency domain. The rolloff in the frequency domain plots is a result of the 2.0 MHz 3 dB bandwidth of the final filter in the RF chain. The raw GPS-SPS data contains the CDMA broadcasts of 5 visible satellites. This same front end configuration allowed enough bandwidth for data capture of 2 of the GLONASS channels. GLONASS data was collected by adjusting the LO of the front end to translate channels 21 & 22 to the resulting sampled bandwidth.

Acquisition is the search for the parameters necessary to identify the signal and begin tracking. In the case of GPS-SPS, this includes the signal's spreading (Coarse/Acquisition (C/A)) code, carrier frequency, and code

phase. GLONASS reduces the search space by one parameter as it uses the same spreading sequence on each frequency channel. The search can be visualized as a matrix (2-D for GLONASS and 3-D for GPS) where every entry must be tested until one is found corresponding to the correct set of parameters. This search space must be bounded with a defined step size. For GPS-SPS there are 32 possible C/A codes. The possible carrier frequency, which differs as a result of Doppler, is bounded for most users to  $\pm 10$  kHz from nominal and is searched in 500 Hz bins. Lastly the spreading code is 1023 chips for GPS-SPS and 511 chips for GLONASS and is searched in  $\frac{1}{2}$  and  $\frac{1}{4}$  chip increments, respectively, over a single code period.

There are a number of popular spread spectrum signal acquisition algorithms [8]. However, most commercial GPS-SPS receivers tend to use the serial search technique. The popularity of this technique is most likely due to the fact that the digital correlator/accumulator hardware can be used not only for tracking, but also acquisition if serial search is employed. In order to demonstrate the flexibility of the software radio approach, multiple acquisition algorithms have been coded.

The postprocessing approach allowed for a more comprehensive evaluation of each of the acquisition techniques. First, the code phase search was stepped in terms of samples. Second, the search conducted was exhaustive, that is every point in the search space was evaluated. In a traditional receiver, points in the search space are sequentially tested until a threshold is crossed indicating a potential match has been obtained and control is transferred to attempt tracking. In serial search there are well-defined equations to calculate the threshold that also determines the associated probabilities (missed detection and false acquisition) [9].

In the standard serial search routine, the signal is converted to baseband using a frequency entry from the test matrix and multiplied by the spreading code with a code phase entry from the test matrix. The resulting data points are accumulated over a single code period and that measurement is used to determine if the correct entry from the matrix has been found. Although this is a well-established technique, the disadvantage is that all test points in the matrix must be evaluated serially, as implied by the name. From the earlier discussion exhaustive testing of single C/A code or GLONASS frequency will require evaluating:

$$\left( \frac{10000}{500} \right) \left( \frac{1023}{\frac{1}{2}} \right) \approx \left( \frac{10000}{500} \right) \left( \frac{511}{\frac{1}{4}} \right) \approx 40920 \quad (1)$$

possible entries in the test matrix. This search space can often be reduced through knowledge of: the satellite almanac data and current user position and time estimates. This will indicate which satellites/frequencies should be tested first as an attempt to reduce acquisition times, but will not improve an exhaustive search as it does not eliminate any of the search space rather it provides a good initial estimate. This technique has been implemented and tested successfully in software. Results on the collected data set will be presented following the discussion of all applied acquisition methods.

Two published improvements on the serial search technique have been implemented for use with the software radio. The first parallelizes the frequency search space [10]. In this case the raw data is multiplied by the spreading code with a code phase from the test matrix, then the Fourier transform of the resulting data set is taken.



All possible frequency bins are checked for the resulting carrier modulated only with the navigation data which, if found, would indicate the proper code phase had been utilized and the bin in which it was located would provide the necessary frequency information. There is no longer a need to search the various frequency entries of the test matrix, however, the computational requirements of the Fourier transform is substituted for the reduced search space.

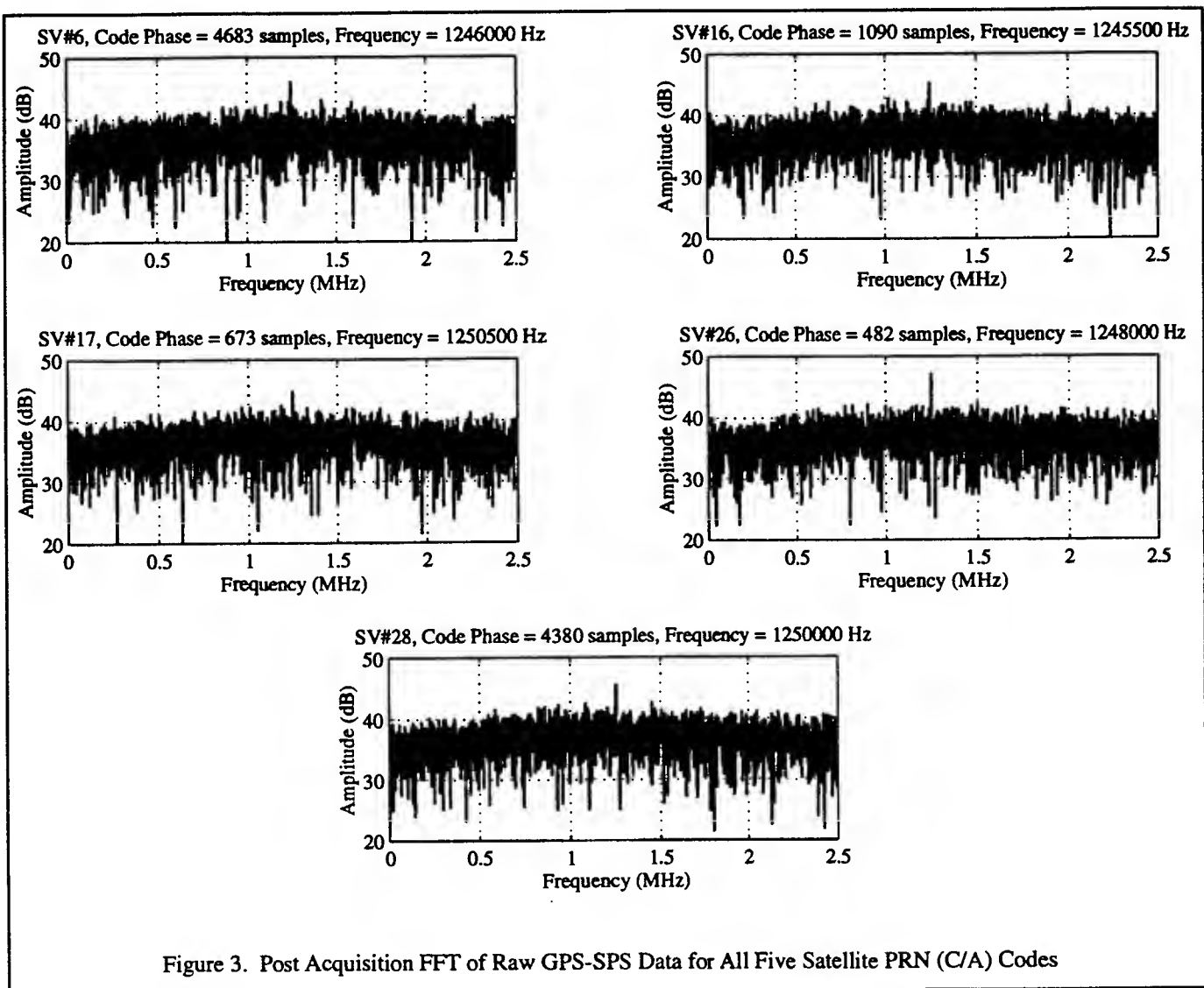
The second technique parallelizes the code phase search also through the use of the Fourier transform [11]. The raw data is converted to in-phase and quadrature baseband components using a frequency entry from the test matrix. The Fourier transform of this data is taken and multiplied by the complex conjugate of the Fourier transform of the spreading code. The inverse Fourier transform is then applied to revert back to the time domain. Since multiplication in the frequency domain acts as convolution in the time domain, the resulting data represents the circular convolution at all possible code phases for that particular frequency. Although this requires the computation of the complex Fourier and inverse Fourier transforms, the search space is reduced to only the possible frequency bins. Since the almanac data can often provide optimal frequency starting points, this technique can significantly decrease acquisition times.

The digital correlator/accumulator, popular in the great majority of receivers, make these acquisition techniques impractical since there is no access to the data prior to accumulation. This limitation illustrates the advantage of the software radio approach. Each of the algorithms were coded and tested with the raw data sets displayed in Figure 2. Although each of the algorithms correctly identified the acquisition parameters for all of the satellites in both data sets, the parallelized code phase search technique greatly reduced the exhaustive search times.

The ability to postprocess the raw data provides interesting plots that give deeper insight into the acquisition process. Figure 3 shows the Fourier transform of the GPS-SPS data depicted in Figure 2 post multiplication with the correct C/A code with the proper code phase. This removes the spreading code and the resulting carrier modulated only with navigation data appears at the appropriate frequency. Figure 4 depicts the

same results for data from both GLONASS frequencies. It is important to note that the acquisition signal processing software implemented is applicable to either GNSS. Figure 5 depicts the exhaustive acquisition search results for all entries of the test matrix for a single C/A code from a visible satellite.

The digital correlator/accumulator, popular in the great majority of receivers, make these acquisition techniques impractical since there is no access to the data prior to accumulation. This limitation illustrates the advantage of the software radio approach. Each of the algorithms were coded and tested with the raw data sets displayed in Figure 2. Although each of the algorithms correctly identified the acquisition parameters for all of the satellites in both data sets, the parallelized code phase search technique greatly reduced the exhaustive search times.



The ability to postprocess the raw data provides interesting plots that give deeper insight into the acquisition process. Figure 3 shows the Fourier transform of the GPS-SPS data depicted in Figure 2 post multiplication with the correct C/A code with the proper code phase. This removes the spreading code and the resulting carrier modulated only with navigation data appears at the appropriate frequency. Figure 4 depicts the same results for data from both GLONASS frequencies. It is important to note that the acquisition signal processing software implemented is applicable to either GNSS. Figure 5 depicts the exhaustive acquisition search results for all entries of the test matrix for a single C/A code from a visible satellite.

#### GNSS SOFTWARE RADIO TRACKING AND DATA PROCESSING

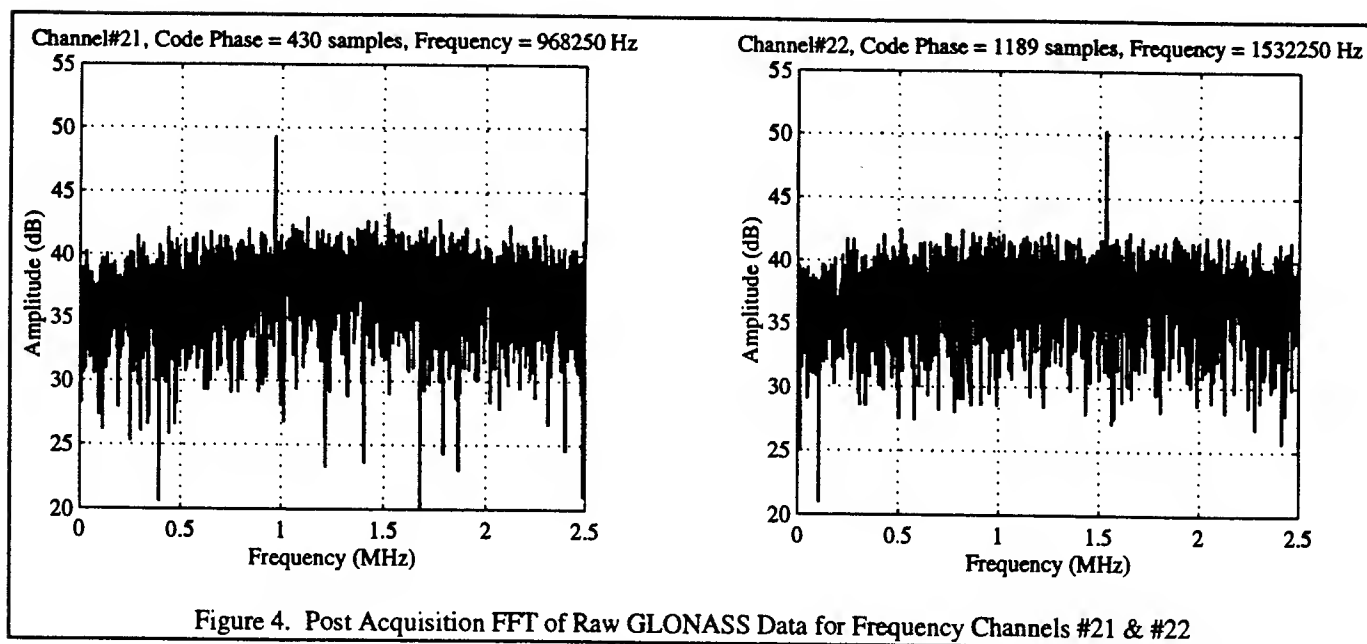
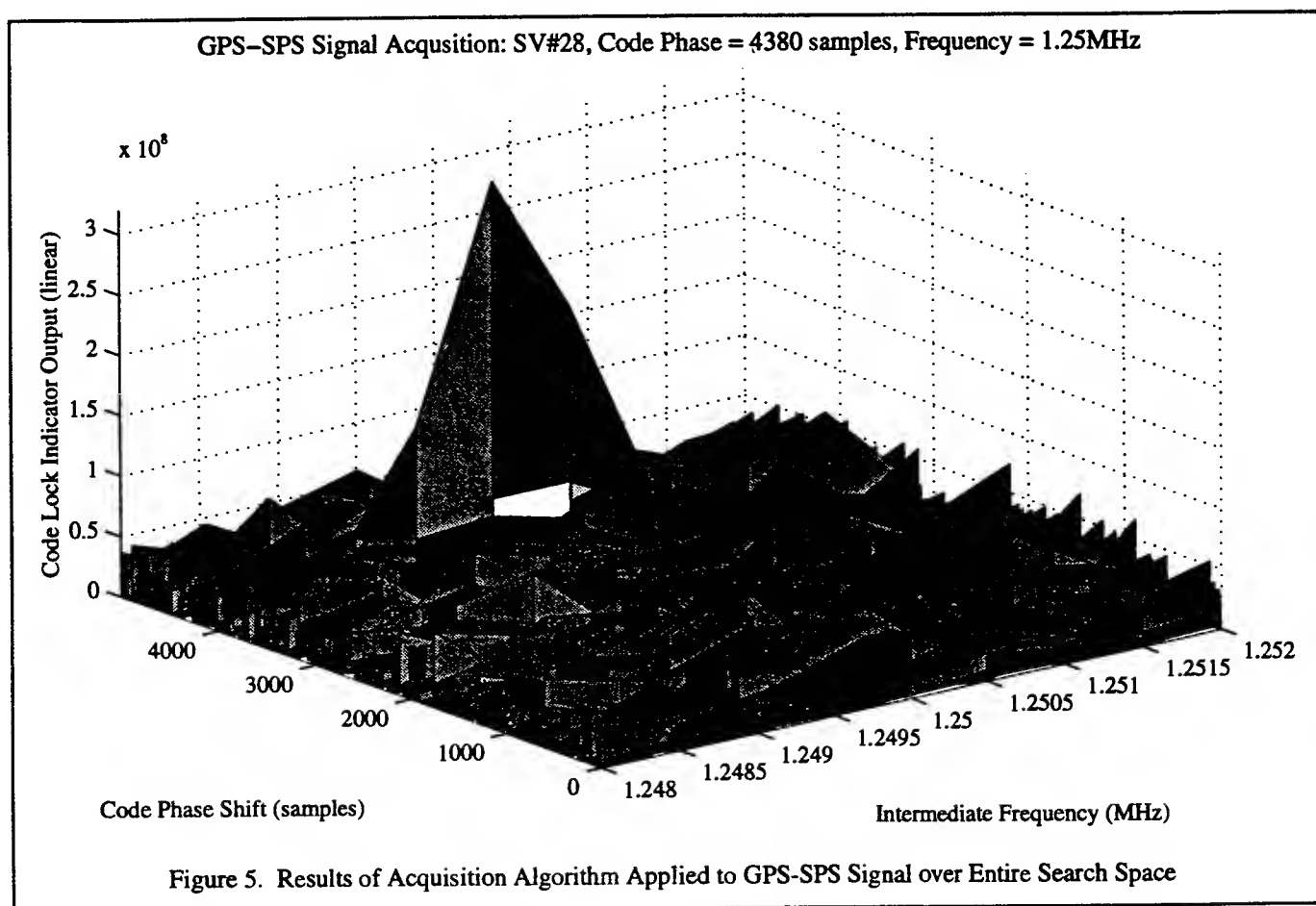


Figure 4. Post Acquisition FFT of Raw GLONASS Data for Frequency Channels #21 & #22

After the acquisition parameters have been identified, the second stage in processing the GNSS signal is tracking and data demodulation. As a participate in the 1996 AFOSR Summer Research Program two distinct approaches were developed and implemented successfully in co-operation with the AAMP group of Wright Laboratories. The goal behind any GNSS tracking algorithm is to precisely align the incoming spreading code with a locally generated version and also to properly decode the navigation data. The combinations of these two tasks allow the calculation of a position solution.

The first technique utilized the traditional tracking loop architecture popular in traditional GNSS receivers. This consists of the two coupled tracking loops, a code tracking loop and a carrier tracking loop. The code tracking loop follows the conventional early-late noncoherent delay lock loop structure [12]. This element seeks to generate a local synchronized version of the incoming spreading code. If the rate of the incoming code changes as a result of the line-of-sight Doppler frequency, the code tracking loop adjusts the locally generated code rate accordingly. This allows the CDMA format of the GNSS signal to be despread and further processed. The carrier tracking loop can be a frequency or phase lock loop. Its purpose is to provide a frequency/phase reference to the code tracking loop and demodulate the navigation data.

The second technique was developed by the AAMP group at Wright Laboratories to track both the code and carrier of a CDMA signal [13]. This technique, known as Block Adjustment of the Synchronizing Signal (BASS) despreads the GNSS signal format and demodulates the navigation data. The code tracking portion of the BASS technique differs from the traditional implementation since the locally generated code rate remains fixed at the nominal code rate. An early and late version of this code is generated and mixed with the incoming signal over 10 code periods. The ratio of the powers in the early and late components provides the additional accuracy needed to adequately track the incoming code. Also this ratio can indicate that the incoming code has slid, as a result of Doppler, more than  $\frac{1}{2}$  of a sample out of synchronization with the locally generated code. When this happens, rather



than attempt to modify the rate of the local code to match the incoming code as is done in the traditional tracking loop, the data set is simply shifted by a single sample in appropriate direction. Using this technique, the locally generated code provides a 'rough' indication of code position (approximate 60 meters using a 5 MHz sampling rate) and the ratio can provide a more precise indication used for the position estimate. Carrier tracking is accomplished in a manner similar to a frequency lock loop. The slope of the resulting baseband signal is used to adjust the carrier frequency. However, navigation bit transitions can result in 180 degree phase shifts. When a navigation bit phase change is detected, it is recorded for later processing of the navigation data, and the phase change is compensated so that the baseband slope can still be used to accurately predict carrier frequency. This slope is based on 10 data points, each of which is determined from 5000 data points (period of the spreading code at 5 MHz).

The data decoded using both of these techniques, in combination with the accurate local code estimates provide the necessary pseudoranges and navigation data parameters to decode a position solution. The GPS-SPS signal specification describes the format of the broadcast navigation data as well as the algorithms to be used in computation of a position solution [14]. These were implemented and applied to the collected data which resulting in a solution within 500 meters of the true position.

## SUMMARY

This paper has presented the initial phase in the development of a GNSS software radio. The advantages of such an implementation to the navigation community were discussed along with potential development obstacles. To date, a software radio front end has been evaluated and various signal acquisition and tracking algorithms, including those which are not applicable in tradition GNSS receiver designs, have been implemented and tested successfully. This has established a position estimate based on software-only processing result of the GPS-SPS signal accurate to within 500 meter.

The next step in the development is the real time implementation. Thus far all results presented have been based on data which was collected and postprocessed. Although this is extremely effective as it allows the necessary software debugging, the ultimate goal is a real time implementation. This is currently under investigation using the Texas Instruments TMS320C80 Digital Signal Processor.

In addition to the real time implementation, the data collection hardware and postprocessing abilities allow a methodical observation of the results of any processing step of each algorithm. This will allow better analysis into the algorithms and should result in processing techniques superior to those used in traditional implementations.

## ACKNOWLEDGMENTS

This work has been sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF and the AAMP group of Wright Laboratories. The material contained within this paper is based on material submitted and approved for publication in the proceeding of the 1996 Institute of Navigation Annual Meeting (ION GPS-96), Kansas City, Missouri, Sept. 17-20, 1996 and also documentation from the 1996 Summer Research Program Final Report.

## REFERENCES

- [1] Vaughan, Rodney G., Scoot, Neil L. and White, D. Rod, "The Theory of Bandpass Sampling", IEEE Transactions on Signal Processing, Vol. 39, No. 9, September 1991, pp. 1973-1984.
- [2] Akos, Dennis M., and Tsui, James B. Y., "Design and Implementation of a Direct Digitization GPS Receiver Front End," IEEE Transactions on Microwave Theory and Techniques, Dec. 1996.
- [3] Mitola, Joe, "The Software Radio Architecture," IEEE Communications Magazine, May 1995, Vol. 33, No. 5, pp. 26-38.
- [4] Brown, Alison and Wolt, Barry, "Digital L-Band Receiver Architecture with Direct RF Sampling," IEEE 1994 Position Location and Navigation Symposium, Las Vegas, Nevada, April 11-15, 1994, pp. 209-215.
- [5] Wepman Jeffery A., "Analog-to-Digital Converters and Their Application in Radio Receivers," IEEE Communications Magazine, May 1995, Vol. 33, No. 5, pp. 39-45.
- [6] Baines, Rupert, "The DSP Bottleneck," IEEE Communications Magazine, May 1995, Vol. 33, No. 5, pp. 46-55.

- [7] Akos, Dennis M., and Braasch, Michael S., "A Software Radio Approach to Global Navigation Satellite System Receiver Design," 1996 Institute of Navigation Annual Meeting, Cambridge, MA, June 1996
- [8] Rappaport, Stephan S., and Grieco, Donald M., "Spread-Spectrum Signal Acquisition: Methods and Technology," IEEE Communications Magazine, June 1984, Vol. 22, No. 6, pp. 6-21.
- [9] Holmes, Jack K., and Chen, Chang C., "Acquisition Time Performance of PN Spread-Spectrum Systems," IEEE Transactions on Communications, Aug. 1977, Vol. 25, No. 8, pp. 778-783.
- [10] Cheng, Unjeng, Hurd, William J. and Statman, Joseph I., "Spread-Spectrum Code Acquisition in the Presence of Doppler Shift and Data Modulation," IEEE Transactions on Communications, Vol. 38, No. 2, February 1990, pp. 241-250.
- [11] van Nee, D. J. R. and Coenen, A. J. R. M., "New Fast GPS Code-Acquisition Technique Using FFT," Electronic Letters, January 17, 1991, Vol. 27, No. 2, pp. 158-160.
- [12] Ziemer, Rodger E. and Peterson, Rodger L., "Digital Communications and Spread Spectrum Systems," Macmillian Publishing Co., 1985.
- [13] Tsui, J. B. Y., Akos, Dennis, and Stockmaster, Michael, "BASS: Block Adjustment of Synchronizing Signal," Wright Laboratories, Patent Pending.
- [14] US Department of Transportation, Global Positioning System Standard Positioning Service Signal Specification, Second Edition, June 2, 1995.

EXPERIMENTAL AND NUMERICAL STUDY OF SHEAR LOCALIZATION AS AN INITIATION  
MECHANISM IN ENERGETIC SOLIDS

Richard J. Caspar  
Graduate Research Assistant  
Professors James J. Mason and Joseph M. Powers  
Faculty Advisors  
Department of Aerospace and Mechanical Engineering

University of Notre Dame  
365 Fitzpatrick Hall  
Notre Dame, IN 46556-5637

Final Report for:  
Summer Research Extension Program  
Wright Laboratories

Sponsored by:  
Air Force Office of Scientific Research  
Bolling AFB, Washington DC

and

Wright Laboratories

July 1996

# EXPERIMENTAL AND NUMERICAL STUDY OF SHEAR LOCALIZATION AS AN INITIATION MECHANISM IN ENERGETIC SOLIDS

Richard J. Caspar  
Graduate Research Assistant  
Department of Aerospace and Mechanical Engineering  
University of Notre Dame

## Abstract

This thesis considers the behavior of energetic and inert solids subjected to simple shear loading. Data from a torsional split-Hopkinson bar, built for this study, was reduced to determine shear stress and shear strain characteristics of these materials. These results were then used to calibrate a constitutive law for stress, including the effects of strain and strain rate hardening and thermal softening. A one dimensional finite difference study of shear localization was performed, modeling the effects of thermal conductivity, viscoplastic heating and Arrhenius kinetics. Results revealed shear localization and reaction initiation in the explosives simulated. Experimental failure of the inert solids, however, occurred at shear strains significantly lower than those predicted by theory. This has been attributed to the presence of failure mechanisms other than shear localization, which were not included in the theoretical model. It is concluded that the tested energetic materials are not expected to shear localize or initiate under the conditions considered.



# EXPERIMENTAL AND NUMERICAL STUDY OF SHEAR LOCALIZATION AS AN INITIATION MECHANISM IN ENERGETIC SOLIDS

Richard J. Caspar

## 1 Introduction

This report will address, experimentally and theoretically, the behavior of various metals, solid explosive simulants, and solid explosives subject to simple shear loading. In addition, this report will consider reaction initiation in the energetic materials as a result of a mechanism known as shear localization or shear banding. In this section, a description and review of the pertinent work is given. Two softening mechanisms which lead to shear banding will be discussed: viscoplastic thermal softening and void nucleation and growth. Pertinent work performed in the study of shear localization in explosives is also discussed.

### 1.1 Overview

The motivation for this report lies in the development of insensitive munitions, which are resistant to accidental detonation. Insensitive munitions are desired for many reasons. First, insensitivity lessens safety risks in the storage and handling of these devices. In addition, it is desired to prevent sympathetic detonation, in which the detonation of one device causes others to detonate. Another motivation for this report comes in the field of deep earth penetrators. These devices are designed to travel through tens of feet of rock, concrete and earth; hence, a significant amount of deformation is inherent within the penetrator. It is thus desired to design these munitions to be insensitive to this deformation.

In order to develop insensitive munitions, it is necessary to more fully understand the behavior of explosives. As full scale tests on explosives are often costly and time consuming, it is desirable to develop computer models and simple bench-top experiments which predict the deformation and initiation of these materials. There are numerous finite-element packages, such as EPIC and ABAQUS, which have been designed to predict material deformation. To date, limited data exists to develop constitutive models for explosives to use as input into these packages. One of the foci of this study is thus to determine the material properties of various explosive simulants and explosives. These properties are determined through the use of an experimental apparatus known as the torsional split Hopkinson bar (TSHB), which was built by this author. Results obtained using this apparatus reveal shear stress-shear strain properties at strain rates of  $10^2$  to  $10^4 \text{ s}^{-1}$ , for tested materials. In addition, photographs of the deformation are taken with an ultra-high speed camera, capable of framing at a rate of 2 million frames per second, to observe failure mechanisms.

The TSHB has previously been used to determine material characteristics for metals, in which failure often occurs due to a mechanism known as shear localization. Shear localization is also known to be one of the initiation mechanisms in solid explosives [Field *et al.*, 1982], also, it is one of the least understood mechanisms. Much of the studies on initiation, however, have been performed under shock and impact conditions, in which the stresses within the explosives are much greater than the yield stress, thus making the effect of the strength of the materials insignificant [Frey (1981), Boyle *et al.* (1989), Chou *et al.* (1991)]. As a result, little is known about the sensitivity of explosives under lower stress deformations, where the strength of the material becomes significant. In deep earth penetrators, it is surmised that explosives undergo significant deformation at high strain rates and relatively low stresses (on the order of the yield stress of the material) and low pressures, in which the material strength is thought to affect the deformation of explosives. One of the detonation mechanism expected to dominate under such conditions is shear localization. An additional goal of the experimental tests is thus an attempt at observing shear localization in explosives deforming in simple shear.

Figure 1 describes the mechanism of shear localization. In Figure 1a, a portion of an undeformed material is sketched with thin lines inscribed on its surface. When this material is sheared, the scribe lines begin to slant at a uniform angle, as seen in Figure 1b. This form of deformation is known as homogeneous deformation. Increased straining into the plastic range results in hardening of the material. In addition,

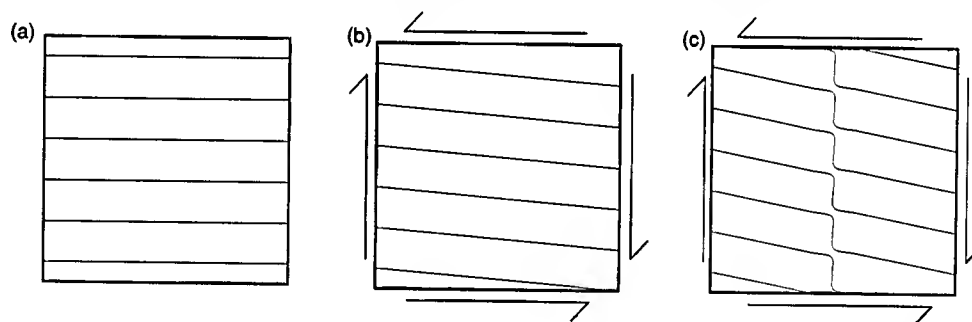


Figure 1: Schematic of the shear localization process. (a) Undeformed grid lines, (b) Homogeneous deformation, (c) Shear localization

if there is a geometric discontinuity, void, scratch or some other material weakness, straining near that discontinuity will occur at a higher strain rate, which also hardens the material. This increased local deformation, however, also causes plastic heating of the material. If the straining occurs at high strain rates (typically greater than  $10^2 \text{ s}^{-1}$ ), there is not enough time for the generated heat to be conducted away. The local increase in heat results in thermal softening of the material. If this process dominates over the hardening due to strain and strain rate effects, the material strength decreases. As a result of this local softening of the material, deformation is localized into a thin planar region, as depicted by the scribe line deformation of Figure 1c. This final process is known as shear localization or shear banding.

In addition to the experimental tests that were performed in this report, a numerical model was developed to predict the deformation of a material in simple shear. The governing equations of conservation of momentum and energy are used in this model. In the discussion of shear localization in the following section, two softening mechanisms will be discussed which are known to lead to shear banding: thermal softening and microvoid nucleation and growth. In this report, a constitutive law for the shear stress is utilized, in which the effects of strain and strain rate hardening, and thermal softening are included; microvoid nucleation is neglected. Since it has been shown that the choice of a particular constitutive law does not significantly effect the results [Wright (1987), and Batra and Kim (1991)], a simplified power law will be implemented. Also included in the model is the effect of thermal conductivity due to its importance in achieving accurate temperature predictions [Batra and Kim, 1991]. Finally, exothermic reaction is modeled by an Arrhenius kinetic law. Despite results attesting to the fact that localization is a multidimensional process [Marchand and Duffy (1988) and Giovanola (1988 a,b)], a one dimensional model will be developed, since this model is sufficient in yielding important information about the shear band temperature profile in solid explosives. The pertinence of these effects is discussed in the following sections.

The novelty of this report first lies in the testing of explosive simulants with the torsional split-Hopkinson bar and the determination of their shear stress and shear strain characteristics. Also, the implementation of Arrhenius kinetics in the study of simple shear deformation of explosives is new. In addition, most researchers studying simple shear deformation have used a finite-element formulation for solving the governing equations. In this report, the equations are solved by a finite-difference method.

## 1.2 General Reviews

This section presents a brief review of works on the high strain rate behavior of materials as well as the initiation mechanisms in solid explosives. A compilation of works on shock wave and high strain rate phenomena in materials is presented by Meyers *et al.* (1992). Meyers (1994) also discusses these dynamic events in materials. These books discuss various failure mechanisms occurring at high strain rate in materials, including shear localization. In addition, works studying high strain rate effects in explosives are discussed.

Bowden and Yoffe (1985) performed an extensive review of experimental works on explosive mechanics in order to categorize the various mechanisms of initiation in solid explosives. They concluded that initiation could occur by the adiabatic compression of small entrapped gas bubbles; the formation of hot spots on confining surfaces, extraneous grit particles and intercrystalline friction of explosive particles; and the viscous heating of rapidly flowing explosive as it escapes impacting surfaces. Many authors, however, have discounted gas compression as the controlling mechanism for hot spot formation [Frey, 1985 and Kang *et al.*, 1992].

Field *et al.* (1982) performed impact tests on thin layers of several explosives, reporting photographic evidence for the formation of initiation due to many of the previously stated mechanisms, as well as some

additional mechanisms. In these tests, the explosives were subject to stresses significantly larger than the yield stresses of the materials. They concluded that the following mechanisms contributed to ignition of explosives: adiabatic shear banding, adiabatic heating of gas spaces, viscous flow, frictional rubbing, hot spots at crack tips and triboluminescence. The authors found that viscous heating could play an important role in liquids, but could only lead to significant heating in solids when considered in conjunction with the other mechanisms listed. They also determined that the propagation of cracks alone would not lead to ignition. Instead, they proposed that fracture of an explosive crystal would produce a gaseous void which could in turn lead to adiabatic heating and hot spot generation. In conclusion, the authors note that no one mechanism is the dominant means of ignition, and that small changes in the experimental conditions can lead to the formation of hot spots due to different mechanisms. For detailed reference on detonation theory with discussion of experiments, see Fickett and Davis (1979).

### **1.3 Adiabatic Shear Banding in Metals**

Shear banding, as an initiation mechanism in explosives, is only simply understood. It is known that metals subject to high strain rate loading in association with high speed machining, cutting and forming, as well as in impact and penetration, often experience highly localized plastic deformation, known as shear bands. The thickness of these shear bands is typically on the order of micrometers, and they have been known to develop in times on the order of microseconds [Marchand and Duffy (1988), Giovanola (1988 a,b)]. Due to the significant amount of localized viscoplastic work on such a short time scale, highly localized temperatures of about 1000°C are observed. Although formation of these shear bands is typically followed by fracture, failure in any accepted sense of the word occurs with formation of the shear band since the material has lost its load carrying capacity [Marchand and Duffy, 1988]. Hence, a significant amount of research has been performed on shear banding as a failure mechanism in structural materials.

#### **1.3.1 Experimental Observations**

It is generally understood that adiabatic shear localization begins because thermal softening dominates over strain and strain rate hardening in the deformation process. When a material is plastically strained, dislocations begin to slip within the material, accumulating at grain boundaries. As these dislocations coalesce, there becomes less room for dislocations to slip, thus resulting in strain hardening of the material [Lubliner, 1990]. If the rate at which this straining increases, viscous stress will further resist deformation, which contributes to a process known as strain rate hardening. An important result of dislocation slip is the generation of heat. Rogers (1979) has concluded that approximately 90% of this plastic work is converted into heat, while the remaining is stored in the generation and arrangement of dislocations. This increase in heat tends to free the motion of dislocations, resulting in thermal softening of the material. The generation of heat

then results in further plastic strain, which causes further increases in temperature. At high enough strain rates, there is not enough time for the heat generated to be conducted away; the deformation thus becomes adiabatic. If a material is experiencing high strain rate adiabatic deformation, and there is a material or geometric weakness, such as a void or scratch, straining will increase locally, causing an instability. If the properties of a given material are such that the mechanism of thermal softening dominates over strain and strain rate hardening, deformation will localize into a planar region, known as a shear band.

Zener and Hollomon (1944) were among the first to describe the process of shear localization in detail. They state that a necessary condition for shear localization is when a maximum in the homogeneous, adiabatic stress strain curve exists, beyond which deformation cannot be homogeneous and strength decreases with increasing strain. When the strain at this maximum is surpassed, an instability will then arise, in which a region deforms at a greater rate than the surrounding material, causing it to weaken and further strain, while the surrounding material is no longer strained.

In order to study the deformation and temperature distribution across a shear band, Marchand and Duffy (1988) performed tests in simple shear on thin tubular specimens, by means of a torsional split-Hopkinson Bar. Previous applications of this device only resulted in average values of shear stress and shear strain. Marchand and Duffy were of the first researchers to perform detailed measurements of the shear band formation. They used high speed photographs to study the deformation of fine lines etched on the specimen's surface in order to determine the local shear strain. From an analysis of their photographs, they concluded that shear banding occurs in three distinct stages. In the first stage, the material is undergoing homogeneous deformation, in which the grid lines are inclined at a constant angle. In the second stage, the material undergoes inhomogeneous deformation, in which the etched lines are curved. As deformation continues in this stage, there is a continuous increase in the localized strain, while the width of the inhomogeneity decreases. It is important to note that, in this stage, the deformation remains uniform in the circumferential direction. The decrease in the stress, however, is never large over this region. The third stage begins at the value of nominal strain where the stress first starts to drop severely. The deformation becomes severely localized in a thin plane. It is at this time that the one dimensional assumption of localization breaks down. Marchand and Duffy observed that the axial position of the maximum local strain is not the same for all points in the circumferential direction indicating that the shear band originates in several locations or that it originates in one location and propagates around the circumference of the specimen.

The maximum shear strain reached in the shear band, with an applied shear strain rate of  $1600 \text{ s}^{-1}$ , is 1900%, with a shear band width found to be  $20 \text{ }\mu\text{m}$ . These large strains demonstrate the weakening effect of shear banding. For dynamic loadings, deformation is often localized to a small region, which is forced to absorb the majority of the deformation. As a result, the strength of the whole specimen is not utilized, causing failure at much lower nominal strains than in quasi-static experiments.

Marchand and Duffy also performed temperature measurements across the shear band with infra-red radiation detectors. Results for an applied strain rate of  $1400\text{ s}^{-1}$  revealed a temperature spike at the location of maximum shear, with a maximum recorded temperature of  $590^{\circ}\text{C}$ . This value represents an average temperature over a region which is greater than the width of the shear band; from a calculation taking into account the width of the shear band, temperatures as high as  $1000^{\circ}\text{C}$  are surmised.

In a similar study, Giovanola (1988,a) performed tests on VAR 4340 steel. He used high speed photography to observe the deformation and infrared detection to determine temperature measurements. In accord with Marchand and Duffy, Giovanola determined a maximum shear strain at failure of 2000%. It is also pertinent to note that Giovanola observed inhomogeneous deformation prior to shear banding, as was observed by Marchand and Duffy (1988). In a companion paper, Giovanola (1988,b) performs fractographic and metallographic observations of the failure surface to determine that shear banding is a result of thermo-plastic instability and microvoid nucleation and growth. In addition, Giovanola notes that failure occurred on a number of parallel planes connected by well defined steps. This observation supports the claim that shear bands nucleate at a number of locations in the softened region and propagate around the specimen.

### 1.3.2 Theoretical Predictions

As a result of the significant amount of experimental data related to dynamic simple shear, determined from the torsional Hopkinson bar, and due to the mathematical ease in its modeling, much of the analytical work in the study of shear localization has been performed in connection with thermoviscoplastic simple shear deformation. In order to theoretically model the problem, governing equations are developed to model the relevant physical conservation principles. These principles do not form a complete set and are hence supplemented by constitutive equations, through which specific materials are modeled. A review of past works in the field of adiabatic shear localization is presented by Rogers (1979).

In the development of a constitutive model for stress, there are two schools of thought as to the method by which the strength is softened. The first theory is that the stress is reduced as a result of heat being generated from viscoplastic deformation. This process results in thermal softening, which dominates over strain and strain rate hardening. Alternatively, many researchers have studied the softening of the stress as a result of microvoid nucleation and growth. This theory is typically formulated to state that at some critical strain, voids begin to nucleate in the material, thus reducing the cross sectional area, and hence the strength of the material. Microvoid nucleation and growth is generally presented in conjunction with thermal softening, microvoid nucleation being the trigger for thermal softening. Meyer (1992) gives a review of some of the constitutive relations which have been used for high strain rate applications. In the present report, a numerical model will be developed which takes into account softening due to viscoplastic work alone; microvoid formation will be neglected.

In existing studies of thermal softening induced shear localization, there is a significant amount of discrepancy in the choice of the constitutive equation. In order to address this issue, Wright (1987) compared the results of four commonly used viscoplastic constitutive relations. The constitutive laws used were 1) an Arrhenius stress law, 2) the Bodner-Partom-Merzer law, 3) a simple power law, and 4) the Litonski law, which were all calibrated over the same data. He found that the results were both qualitatively and quantitatively similar, with the results within 5% of each other for shear strain rates up to  $10^4 \text{ s}^{-1}$ , a value far in excess of the calibration conditions. For even larger values of shear strain rate and high temperatures, the Bodner-Partom-Merzer law begins to diverge from the other solutions. Wright states that since there is a significant difference between the strain rate and temperature at the center of the shear band from those of the calibration conditions, the actual structure of a real shear band would be expected to be somewhat different from that predicted by any given constitutive law. He thus thought it surprising that the trends predicted by the constitutive laws were as similar as found.

Another source of discrepancy in previous researchers is in the role of thermal conductivity. To clarify this issue, Batra and Kim (1991) compared the results of three different constitutive relations, while varying the thermal conductivity. In this study, the researchers considered a thermoviscoplastic block undergoing one dimensional simple shearing deformations, with strain and strain rate hardening and thermal softening. The thickness of the block was taken to vary smoothly with a 5% decrease in thickness at the center. The constitutive laws considered were the Litonski law, the Bodner-Partom law, and the Johnson-Cook law. Batra and Kim found that the results from the three constitutive laws are extremely similar, both qualitatively and quantitatively, verifying the results achieved by Wright (1987). As a result of these studies, we have determined that the use of a simple power law will be adequate in the study of high speed deformation.

Batra and Kim also reported that large increases in the value of the thermal conductivity delay the initiation of stress collapse and slow down the development of the shear band. However, for realistic values of thermal conductivity there is little effect on the value of nominal strain at which stress collapses, and hence localization occurs. In contrast, the authors found that the rate of evolution of the temperature at the center of the specimen decreases with increasing thermal conductivity, resulting in significant changes in temperature for realistic values of thermal conductivity. It is therefore concluded that thermal conductivity can be neglected when one is only considered with the timing of stress collapse, but it proves to be crucial when considering shear band temperatures. Since this current report is studying thermal reaction initiation in explosives, conductivity will play an important role.

Motivated by this study, Batra *et al.* (1995) performed a thermoviscoplastic analysis neglecting thermal conductivity in order to rank twelve materials according to the critical strain necessary to reach localization in tubular specimens loaded in simple shear. A finite-element method with the Johnson-Cook constitutive law was used in these calculations. The thickness of the tube was taken to vary smoothly with a 10%

decrease at the center. Batra *et al.* assumed that the shear band initiates when there is a catastrophic drop in torque, and ranked the materials according to the corresponding nominal shear strain. They found shear bands to initiate in the following order: tungsten, S-7 tool steel, depleted uranium, 2024-T351 aluminum, 7039 aluminum, 4340 steel, armco iron, carpenter electric iron, 1006 steel, cartridge brass, nickel 200 and OFHC copper, when tested with a shear strain rate of  $5000 \text{ s}^{-1}$ . It is relevant to note that the critical strain was dependent on the size of the initial defect as well as the finite element mesh used. The relative ranking of the materials, however, was independent of these parameters.

In order to study the development of a shear band, Batra and Kim (1992) solved a nonlinear system of equations for a thermoviscoplastic block with the Johnson-Cook constitutive law, including strain and strain rate hardening, thermal softening and thermal conductivity. They used a continuous variation in the thickness to instigate localization. The authors found the same three stage localization process as Marchand and Duffy (1988) with transition to stage two occurring at the time the stress reached its maximum value, and stage three occurring much later, when the stress has dropped to about 90-95% of its maximum value. Batra and Kim also observed the effects of varying the thickness of the block. They found that the defect size has a stronger influence on ductile materials than on less ductile materials, but in all cases, it has a significant effect on the critical strain to reach localization.

In a similar study, Clifton *et al.* (1984) used a simple power law including strain and strain rate hardening and thermal softening but neglected thermal conductivity to perform a study on the critical conditions for shear band formation. They concluded that the primary factor affecting initiation of the shear band is the strain hardening, whereas the strain rate hardening is the primary factor affecting the rate of growth of the shear band. Supplementing these observations, Wright and Batra (1985) used the Litonski flow law to explore the critical strain at collapse. In accord with Clifton *et al.* (1984), Wright and Batra determined that strain rate plays little role on the critical strain at localization. However, they found that the size of an initial temperature perturbation plays a significant role, with the larger perturbation causing localization to occur at a smaller critical strain.

## 1.4 Detonation Mechanisms in Explosives

As was stated previously, shear localization is understood to be one of the mechanism which can lead to ignition in solid explosives subject to high strain rates and relatively low pressures. Traditional studies of shear localization in explosives have been performed in conjunction with high pressures, representing the conditions undergone in shock and impact loading. Frey (1981) developed a model to describe heating in high explosives due to shear banding. He used a linear viscoplastic constitutive law, neglecting the effects of strain rate hardening. The strength was decreased over a  $30^\circ\text{C}$  range after the melting point, the melting point increased linearly with pressure, and the viscosity was dependent on both temperature and pressure.



In addition, Frey used Arrhenius kinetics to model the thermal explosion. This material was then deformed in shear, stimulating localization by setting the strength to zero over a small region within the deformation. Without reaction, Frey's model achieved a maximum temperature which turned out to be independent of strength of the material and thickness of the initial weakness. These parameters did, however, affect the rate of growth of the shear band. The author found that the factors which did affect the temperature were the pressure, strain rate, and viscosity. Frey found that high pressures result in higher temperatures. In addition, under low pressures, it would be very difficult to reach the temperatures in a shear band required to instigate thermal explosion.

In experiments performed on explosives subject to lower pressures and shear deformation, such as in drop weight tests, Boyle, Frey and Blake (1989) confirm the numerical observations of Frey (1981). They find that pressure and velocity do indeed have a strong effect on the initiation of solid explosives. Due to a comparison of explosive materials, they also verify that higher viscosities increase the sensitivity to explosion. In a more recent study, Chou *et al.* (1991) studies two theories for the impact initiation of explosives: shock initiation, a pressure dependent theory, and shear initiation, a temperature dependent theory. Chou *et al.* state that there are three means by which heat can be generated in explosives: shock compression energy, plastic work and viscous work. They state, in accord with previous researchers, that plastic work, in the absence of high pressures, would not generate enough heat to produce thermal explosion, since failure occurs before a significant amount of straining occurs. Chou *et al.* further state that once a material has reached its melting temperature, the effect of plastic work becomes negligible; heating then results from viscous work, which is capable of increasing the temperature well above the melting point. It is known, though, that brittle materials become more ductile under pressure; in fact, Chou *et al.* state that pressure can raise the stress and strain to as much as ten times as high as a material's uniaxial value. This effect thus increases the importance of considering heating by plastic work, when considering a material under hydrostatic pressure.

In order to compare the effect of shock and shear initiation in impacted explosives, Chou *et al.* developed a numerical model similar to that of Frey (1981) and simulated the impact of bare and covered explosives subject to impact. They concluded that for bare explosives impacted by a projectile, shock initiation is dominant and the shear effect is negligible. For covered explosives impacted by a projectile, viscoplastic heating is of importance and shear initiation at the edge of the plug is probable.

## 2 Experimental Method

This chapter discusses an experimental apparatus, known as the torsional split Hopkinson bar (TSHB), which was used to test metals, solid explosive simulants and solid explosives in simple shear. An analysis of the data determined from this apparatus produces average shear stress and shear strain characteristics of the tested material, for a range of shear strain rates ( $10^2 - 10^4 \text{ s}^{-1}$ ). High speed photographs are taken

of the deformation and failure of the specimens, in order to determine their failure mechanism. The data will be used to determine constitutive parameters for input into the numerical model which is presented in the next section. This data can also be used to calibrate the various constitutive laws used in finite element packages such as EPIC and ABAQUS, which can be used in the modeling of explosive mechanics.

The torsional Hopkinson bar is a modification of an apparatus originally discussed by Kolsky (1949, 1953). In his device, thin cylindrical wafer-like specimens were placed between two long elastic bars, aligned along a common axis. The specimen was loaded by propagating a compressive pulse, generated by impacting the bar with a cylindrical projectile of the same material and equal diameter, down one of the bars toward the specimen. A similar device was used by Harding *et al.* (1960) for material testing in tension. The Hopkinson bar was later adapted for tests in torsion, which is discussed by Hartley *et al.* (1985).

There are several reasons why the TSHB is appropriate for the current experiments performed on solid explosives. First of all, in torsional loading, the maximum stress in the specimen occurs on the exterior surface of the material. The largest deformation will thus occur on the exterior surface, making the probability of hot spot formation greatest, where it can easily be observed. Also, the TSHB can be designed to produce a torsional pulse of almost any desired length, and hence, large amounts of deformation in the specimen are possible. In addition, shear is the main form of deformation present in high rate deformation events such as penetration; hence, it is desired to determine stress-strain characteristics in shear, as opposed to compression or tension. Compression and tension test results can be converted into shear data by using a criterion such as the von Mises equivalence relation, but this is not valid for strains above 20% [Hartley and Duffy, 1985]. Further disadvantages of compressive testing result from the Poisson's ratio effect, which causes radial expansion of materials loaded in compression. Additional radial stresses in compressive tests occur due to frictional effects between the specimen and bar. In torsional tests, there is no Poisson's ratio effect, and hence no radial contraction or expansion.

There are, however, some drawbacks to the TSHB. First, tests may only be run on a limited range of strain rates ( $10^2 - 10^4 \text{ s}^{-1}$ ). The lower limit is due to an increase in the noise to signal ratio, while the upper limit is due to the elastic limit of the TSHB. A further drawback is if fracture occurs in the explosive specimen too soon, localization, and hence initiation, is less likely. In addition, the data analysis for this apparatus assumes homogeneous deformation. Hartley *et al.* (1985) have shown that it takes a few reflections of the loading pulse from the ends of the specimen before a state of homogeneous deformation is reached, thus rendering the early time results of the TSHB inaccurate.

## 2.1 Description of Apparatus

The TSHB consists of two elastic cylindrical bars: an incident and transmission bar; a torsional pulley; a clamp; and a specimen. A schematic of this apparatus is included in Figure 2. The two bars are aligned

along a common axis, with a thin walled cylindrical specimen of known geometry joining them. The end of each bar in contact with the specimen is milled to produce a hexagonal socket, into which the specimen or an adaptor is inserted. The adaptor is used for cases in which it is desirable to glue the specimen in place, rather than grip it with the hexagonal socket. The torsional pulley is attached to the end of the incident bar far from the specimen, and the clamp is placed at a variable distance from the pulley, typically several feet.

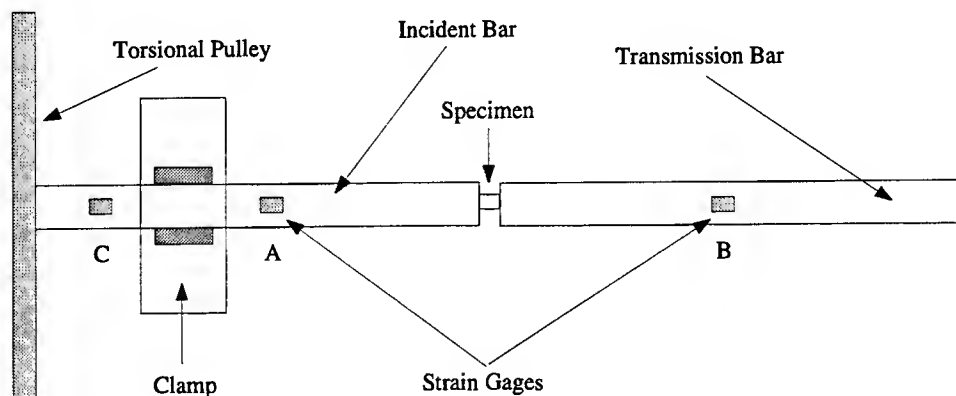


Figure 2: Schematic of the TSHB (not to scale).

The clamp is used to prevent rotation of the incident bar while the torsional pulley is rotated, thus storing a torsional pulse in the bar between the pulley and clamp. The sudden release of the clamp propagates an incident shear strain pulse down the incident bar. The incident pulse reaches the specimen, transmitting some strain through the specimen into the transmission bar and reflecting some back into the incident bar.

The two elastic bars are constructed of 1 in diameter aluminum 7075-T6, 111 in in length. At the ends joining the specimen, a hexagonal socket, of width 0.5625 in and depth 0.25 in, is milled into the bars. Into these sockets, one inserts either a hexagonal specimen or an adaptor machined from 7075-T6 aluminum, to which cylindrical specimens are glued. The hexagonal specimen (see Figure 3) or adaptor is fixed to the bar with 12 set screws, two on each face of the hexagon. The dimensions of the specimens that were used in this study are given in Figure 3. In Figure 3, the central part of the specimen is commonly referred to as

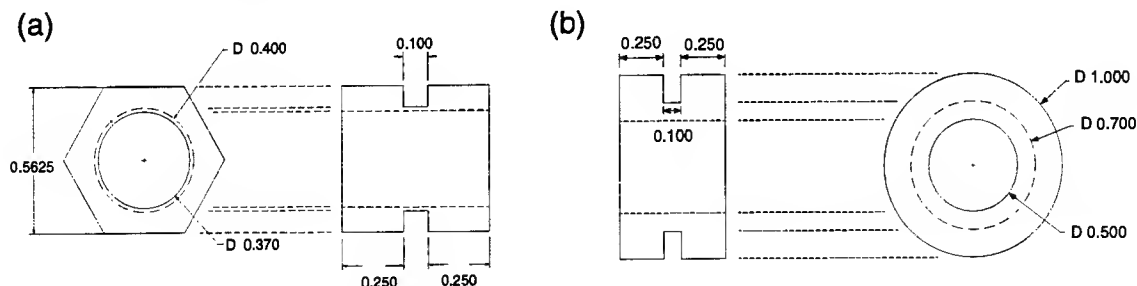


Figure 3: Scaled diagram of the specimens used in the TSHB tests: (a) hexagonal specimen, (b) cylindrical specimen (all dimensions in inches).

the gage length, while the ends are termed flanges. The hexagonal specimen is used to test metals and the cylindrical specimen, due to its ease in machining, is used to test the explosive simulants and explosives. The ratio of the wall thickness of the gage length to the mean diameter of the gage length for the hexagonal and cylindrical specimens are 0.04 and 0.17, respectively. The elastic bars are supported along their length by delrin bearings, which are mounted on adjustable bearing supports. These bearing supports are then fixed to a steel I-beam which supports the whole TSHB apparatus.

A schematic of the torque generating mechanism is included in Figure 4. The torsional pulley is clamped to the incident bar by means of a frictional fit. A hydraulic hand pump is used to pressurize the rams, which lengthen, transferring force into the cable, which in turn rotates the torsional pulley. When the clamp is engaged, this action stores torsional elastic energy in the incident bar.

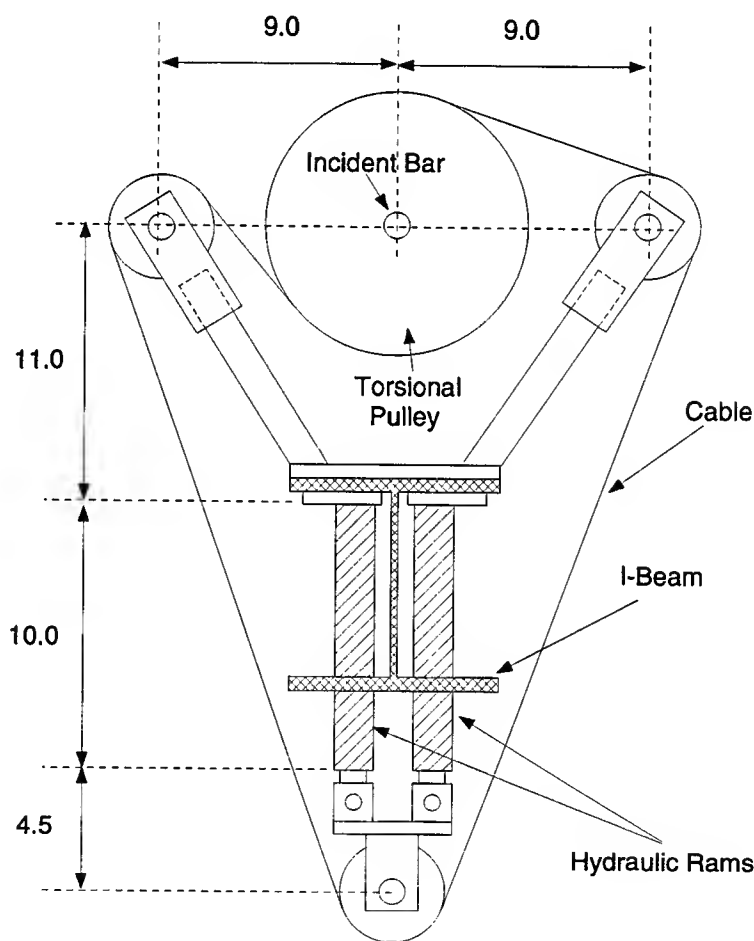


Figure 4: Scaled schematic of the torque generating mechanism (all dimensions in inches).

Integral in the operation of the TSHB is the clamp which stores the torsional pulse. The key to constant strain rate tests is rapid release of the clamp, which propagates an incident torsional pulse towards the

specimen. In order to more fully understand the operation and design of the TSHB, the author spent the summer of 1995 at Eglin AFB in the Advanced Warheads Evaluation Facility (AWEF) making modifications to the design of their TSHB. With the knowledge gained through these efforts, this author modified previous clamp designs resulting in the design shown in Figure 5. The clamp is engaged by pumping a second hydraulic hand pump, which pressurizes a hydraulic C-clamp, which in turn clamps the base of the two clamp faces. This action transmits pressure through the clamp faces onto the Hopkinson bar. In order to release the clamp, the hydraulic pressure is increased until the break element, as seen in Figure 5, fractures, causing the release of the clamp faces. Ideally, the incident pulse would be a square pulse of torsion, with instantaneous rise and fall time and constant magnitude, thus producing deformation in the specimen at a constant strain rate. For optimum functioning of the clamp, it is desired to store large amounts of elastic energy in the clamp in order to achieve quick fracture of the break element and consequently, sudden release of the clamp. This sudden release will produce a pulse with a short rise time and relatively constant magnitude.

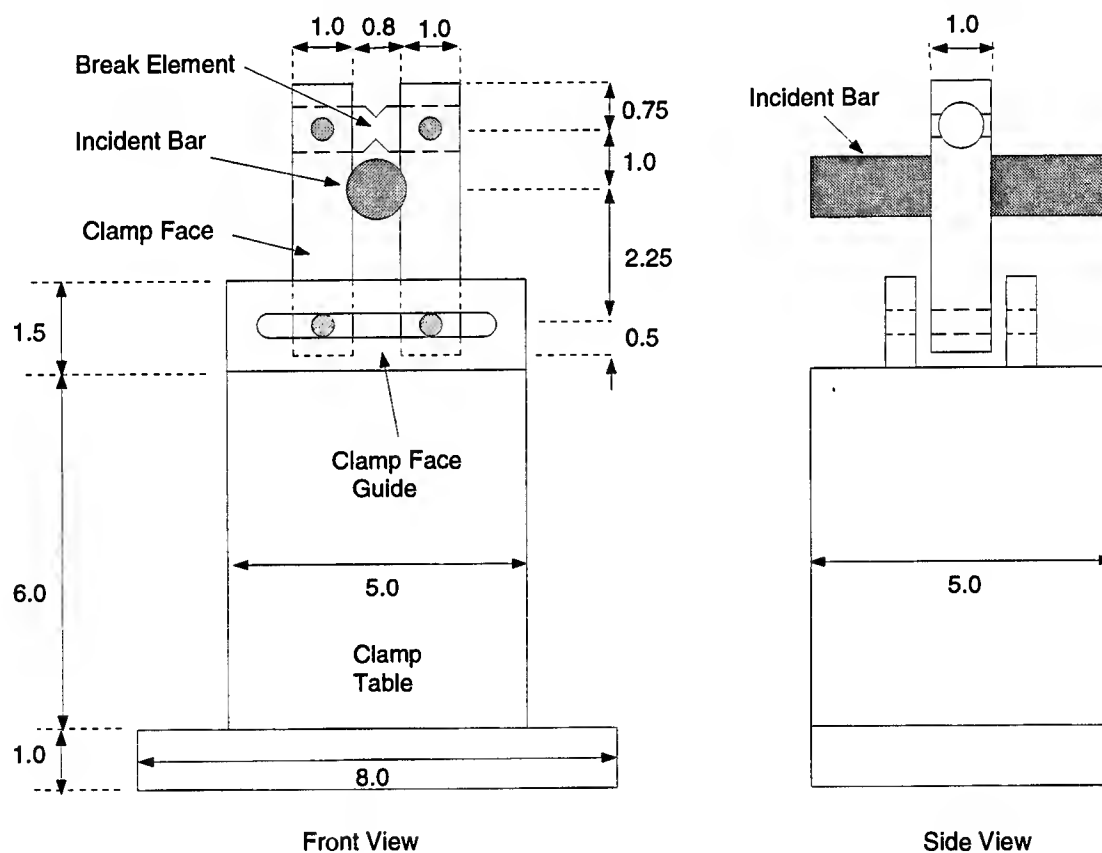


Figure 5: Schematic of the clamp for the TSHB (all dimensions in inches).

In designing a break element, it is desired to use a material with minimum ductility, but not so brittle that it will fail before the clamp is tight enough to store the desired torque. Hartley *et al.* (1985) state that

functional pin materials include aluminum 6061-T6 and 2024-T6. In order to determine the effect of the break element notch geometry on the release of the clamp, the author performed quasi-static uniaxial tension tests with V-notched and square-notched elements. The break element is secured to the clamp faces with a pin, hence, the clamp faces are free to rotate relative to the break element. The element thus experiences almost pure tension, validating the uniaxial notch geometry tests. Results of these tests, reported in Caspar (1996), revealed that the V-notched element performs better than the square notched element.

The break element that was implemented into the clamp design was machined from 0.75 in diameter aluminum 6061-T6 rod, with the diameter at the center of the notch reduced to approximately 0.360 in. The clamp faces are machined from 4340 steel hardened to about 45 on the Rockwell-C scale. In clamping, application of vertical forces that would cause bending and axial pulses, which could result in erroneous data, are avoided by allowing the clamp to move relative to the incident bar. This is accomplished by horizontal slots in the clamp face guides, as seen in Figure 5. In addition, the clamp faces are joined to the guides by pins, allowing the clamp faces to rotate relative to the guide, and hence further eliminating axial and bending pulses by establishing flush contact between the clamp faces and the incident bar.

A typical record of the shear strain pulses recorded with the strain gages at locations A and B (see Figure 2) can be found in Figure 6. The rise time from 10% to 90% of the maximum strain is determined from this data and subsequent tests to be range from 30 – 50  $\mu s$ . In this figure, it can be seen that the reflected pulse, which will be shown to be proportional to the shear strain rate in the specimen, is essentially constant in magnitude while strain is being transmitted. In addition, since the transmitted pulse is shorter than the reflected pulse, the incident pulse is sufficiently long enough to strain the specimen to failure.

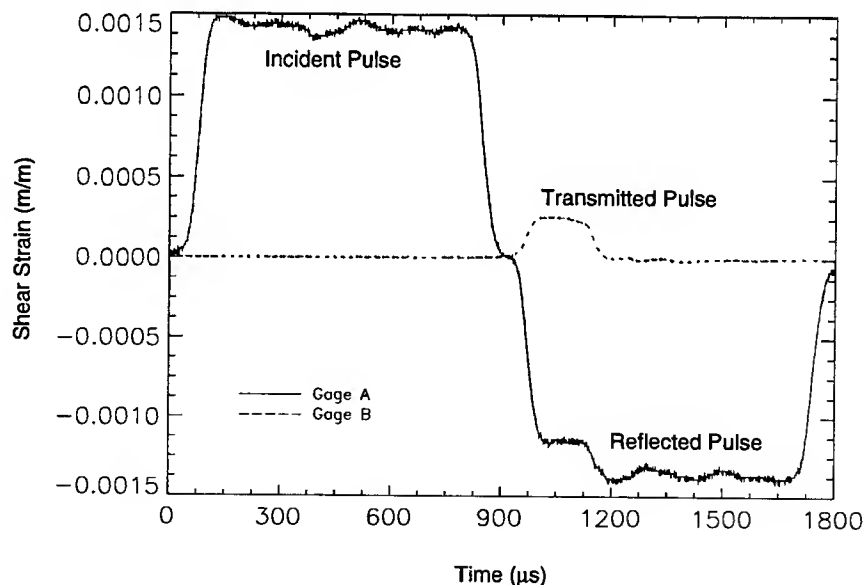


Figure 6: Typical shear strain pulses in a TSHB for 1018 CRS (Test 4).

## 2.2 Analysis

The values of the shear stress, shear strain and shear strain rate experienced in the specimen, when homogeneous deformation in the specimen is assumed, are determined from an analysis of the strain in the incident and transmission bars. In this section, the subscript,  $s$ , is used to denote the properties of the specimen.

Hartley *et al.* (1985) have shown that the shear strain rate in the specimen is proportional to the reflected strain in the incident bar,  $\gamma_R$ . Integration of the reflected strain over time,  $t$ , then provides the shear strain in the specimen,  $\gamma_s$ :

$$\gamma_s(t) = -\frac{2cD_s}{L_s D} \int_0^t \gamma_R(\tilde{t}) d\tilde{t}, \quad (1)$$

where  $c$  is the elastic torsional wave speed in the incident and transmission bars,  $D_s$  is the mean diameter of the specimen,  $L_s$  is the length of the specimen, and  $D$  is the diameter of the incident and transmission bars, and  $\tilde{t}$  is a dummy variable of integration.

Hartley *et al.* have also shown that the transmitted pulse,  $\gamma_T$ , provides a measure of the shear stress in the specimen,  $\tau_s$ :

$$\tau_s = \frac{GD^3}{8D_s^2 t_s} \gamma_T, \quad (2)$$

where  $G$  is the shear modulus, and  $t_s$  is the thickness of the specimen wall.

## 2.3 Data Acquisition and Reduction

The shear strain pulses in the incident and transmission bars are recorded by means of electric resistance strain gage Wheatstone bridges, which are extremely sensitive to small changes in strain. The strain gages are attached at the midpoints of each bar, location A and B in Figure 2. With the gages mounted at the midpoint of the bar, it is possible to record a pulse without its reflection overlapping in time, as long as the pulse is shorter than the length of the bar. Since the clamp is mounted between the torsional pulley and gage station A, the incident pulse, being twice as long as the stored torque, will always be less than the length of the bar. In addition, since gages A and B are each the same distance from the specimen, it is ensured that the reflected and transmitted pulse will commence at approximately the same instant in time. There is also a strain gage bridge mounted 12 *in* from the torsional pulley, gage station C in Figure 2. The purpose for this gage is to record the stored torque, which is used to obtain the expected strain rate incident upon the specimen. Stations A and B consist of four strain gages; one set of 2 torque gages is mounted diametrically opposite another set on the surface of the bar. Each gage is mounted at a 45° angle to the axis of the bar. Strain gages mounted this way will record only shear strain, cancelling any axial and bending strain which may be present. Gage station C consists of one set of strain gages, also mounted at a 45° angle to the axis of the bar.

McConnell and Filey (1993) describe the functioning of a Wheatstone bridge and determine the following equation which calculates the shear strain,  $\gamma_0$ , from the change in the bridge output voltage,  $\Delta E_0$ :

$$\gamma_0 = \frac{2\Delta E_0}{FVG_a N_g} \frac{(1 + \kappa)^2}{\kappa}, \quad (3)$$

where  $F$  is the gage factor,  $V$  is the excitation voltage,  $G_a$  is the amplifier gain,  $N_g$  is the number of active gages in the Wheatstone bridge, and  $\kappa$  is the ratio of resistances in the bridge.

Each bridge is wired to a Measurements Group model 2311 signal conditioning amplifier, which sets the excitation voltage and gain. The amplifier output from gage station A is split, with one lead, as well as the amplifier output from gage station B, sent to a Tektronix model TDS 420 digitizing oscilloscope, which is downloaded to a personal computer. This data is reduced by Equations (3), (1), and (2), to determine nominal shear stress and shear strain characteristics of the specimen. The other lead of the amplifier output from gage station A is sent to a Hewlett Packard model 214A pulse generator. The incident pulse triggers the pulse generator, which sends a second pulse, delayed by some specified time, to trigger a Cordin model 607 light source, which illuminates the deformation of the specimen. Photographs of this process, which are used to determine the failure mechanism of the specimen, are recorded with a Cordin model 330A ultra-high speed camera.

Results from the TSHB were verified by Caspar (1996).

### 3 Model Equations

This chapter introduces the model equations employed in this research. In the first section, the physical problem is described. The assumptions and governing equations are then presented. These equations are scaled and presented in nondimensional form. Finally, the numerical method used to solve the equations is discussed.

#### 3.1 Governing Equations

The current theoretical study of localization is being used to compare with tests performed on the deformation of specimens in the TSHB. Tests have been performed on metals, solid explosive simulants and solid explosives. In the TSHB, a thin walled circular cylinder is dynamically loaded in torsion; the model is thus designed to simulate these conditions. A schematic of the specimen described by the model is included in Figure 7. In this specimen, which has length  $L_s$ ,  $z$  is the axial distance variable,  $\theta$  is the circumferential distance variable, and  $r$  is the radial distance variable. This specimen is loaded by linearly increasing the circumferential velocity,  $v_\theta$ , over time,  $t$ , at  $z = L_s$  to some constant value,  $v_1$ , while holding the velocity fixed at  $z = 0$ . The following assumptions are made in developing the model:



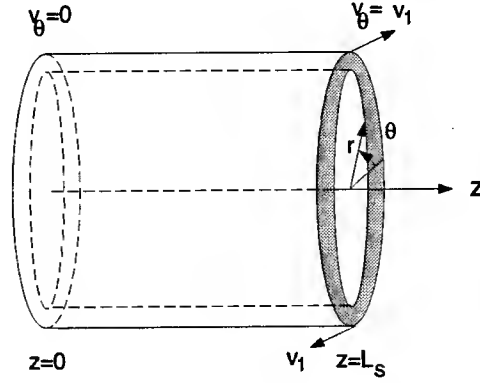


Figure 7: Specimen used for numerical simulation.

- The specimen is initially unreacted, unstressed, and at ambient temperature.
- As we are dealing with a pure torsional problem, there is no component of velocity or displacement in the radial or axial direction:  $v_r = v_z = u_r = u_z = 0$ .
- Due to axisymmetry and the thin walled geometry, there is no variation in the circumferential or radial direction:  $\frac{\partial}{\partial \theta} = \frac{\partial}{\partial r} = 0$ .
- In order to induce localization, we allow the specimen wall thickness to vary with axial position  $z$ . We make the *ad hoc* assumption that this perturbation is sufficiently small so as not to introduce gradients in the radial or circumferential directions. It is noted that alternative methods of perturbation which do not require such *ad hoc* assumptions, such as perturbation in initial velocity, temperature, displacement or strain, could also induce localization.
- In pure torsion, the stress tensor reduces to one component,  $\sigma_{z\theta}$ , the stress on the axial face in the circumferential direction, which will be referred to as the shear stress,  $\tau$ .
- The shear strain is restricted to positive values.
- Plastic deformation is completely converted to heat.
- Heat is only transferred in the axial direction.
- The material undergoes a one-step chemical reaction, with  $A$  denoting the unreacted material, and  $B$  denoting the reacted material.
- The density,  $\rho$ , and the thermal conductivity,  $k$ , are equal in the unreacted and reacted material and, along with the specific heats,  $c_A$  and  $c_B$ , they are constant.

Under these assumptions, the governing equations are stated below:

$$\rho w \frac{\partial v_\theta}{\partial t} = \frac{\partial}{\partial z} (w\tau) , \quad (4)$$

$$\rho w \frac{\partial e}{\partial t} = w\tau \frac{\partial v_\theta}{\partial z} - \frac{\partial}{\partial z} (wq_z) , \quad (5)$$

$$\frac{\partial u_\theta}{\partial z} = \gamma , \quad (6)$$

$$\frac{\partial u_\theta}{\partial t} = v_\theta , \quad (7)$$

$$\frac{\partial \lambda}{\partial t} = Z(1 - \lambda) \exp\left(-\frac{E}{RT}\right) , \quad (8)$$

where  $e$  is the internal energy,  $q_z$  is the heat flux in the axial direction,  $u_\theta$  is the displacement in the circumferential direction,  $\gamma$  is the shear strain,  $\lambda$  is the reaction progress variable, and  $T$  is the temperature. The parameters  $w$ ,  $Z$ ,  $E$ , and  $R$  are, respectively, the thickness of the specimen wall thickness, the kinetic rate constant, the reaction activation energy, and the universal gas constant. Equation (4) models the conservation of linear momentum. Equation (5) models the conservation of energy. Equation (6) is the definition of strain. Equation (7) defines velocity as the time derivative of displacement. Finally, Equation (8) is an Arrhenius kinetics law.

The constitutive equations used in this model are:

$$\tau = \alpha T^\nu \gamma^\eta \left| \frac{\partial \gamma}{\partial t} \right|^{\mu-1} \frac{\partial \gamma}{\partial t} , \quad (9)$$

$$q_z = -k \frac{\partial T}{\partial z} , \quad (10)$$

$$e = m_A e_A + m_B e_B , \quad (11)$$

$$e_A = c_A T + e_A^\circ , \quad (12)$$

$$e_B = c_B T + e_B^\circ , \quad (13)$$

$$m_A = 1 - \lambda , \quad (14)$$

$$m_B = \lambda , \quad (15)$$

where  $\alpha$  is the stress constant; subscripts  $A$  and  $B$  refer to the unreacted and reacted material, respectively;  $e_A$  and  $e_B$  are the internal energies;  $m_A$  and  $m_B$  are the mass fractions;  $c_A$  and  $c_B$  are the specific heats; and  $e_A^\circ$  and  $e_B^\circ$  are the energies of formation. Equation (9) is a constitutive law for stress, proposed by Clifton, *et al.* (1984) where  $\nu$ ,  $\eta$ , and  $\mu$  are the exponents which characterize the thermal softening, the strain and strain rate hardening, respectively. Equation (10) is Fourier's law of heat conduction. Equation (11) is a mixture law. Equations (12) and (13) are the constitutive laws for energy. Lastly, Equations (14) and (15) define the mass fractions.

The following boundary conditions are used in this model:

$$v_\theta(t, 0) = 0 , \quad v_\theta(t, L_s) = \begin{cases} (v_1 - v_0) \frac{t}{t_1} + v_0 & t < t_1 \\ v_1 & t \geq t_1 \end{cases}$$

$$u_\theta(t, 0) = 0, \quad u_\theta(t, L_s) = \begin{cases} (v_1 - v_0) \frac{t^2}{2t_1} + v_0 t & t < t_1 \\ (v_1 - v_0) \frac{t_1}{2} + v_0 t_1 + v_1 (t - t_1) & t \geq t_1 \end{cases} \quad (16)$$

$$\frac{\partial T}{\partial z}(t, 0) = 0, \quad \frac{\partial T}{\partial z}(t, L_s) = 0 \quad t \geq 0.$$

That is,  $v_\theta$  is fixed at one side of the specimen and ramped over a time  $t_1$  from some arbitrarily small velocity,  $v_0$ , to a constant value  $v_1$ . The boundary conditions on displacement are determined by integrating over time the boundary conditions on velocity. Finally, the boundary conditions on temperature are such that the ends of the specimen are insulated. The initial conditions are:

$$v_\theta(0, z) = v_0 \frac{z}{L_s}, \quad u_\theta(0, z) = 0, \quad T(0, z) = T_0, \quad \lambda(0, z) = 0, \quad (17)$$

where the specimen is initially stress free, unreacted and at a uniform temperature,  $T_0$ .

In order to induce localization at the center of the specimen, the thickness of the tube is perturbed so that there is a continuous variation in its thickness with the thinnest portion being at the center, an amount of  $h_p$  less than at the edges. The exact form of this perturbation is as follows:

$$w = w_0 - \frac{h_p}{2} \left[ 1 - \cos \left( \frac{2\pi z}{L_s} \right) \right]. \quad (18)$$

Next, the governing equations are reduced through insertion of the constitutive laws. First, by differentiating Equation (6) with respect to time and Equation (7) with respect to space, and equating the results, one determines the following expression relating shear strain rate with velocity gradient:

$$\frac{\partial \gamma}{\partial t} = \frac{\partial v_\theta}{\partial z}. \quad (19)$$

Now, Equations (6), (9), and (19), are inserted into Equation (4). The energy equation, Equation (5), is reduced by substitution of Equations (8)–(13) and (19). Finally, Equations (7), (8) and (18) are restated:

$$\rho w \frac{\partial v_\theta}{\partial t} = \frac{\partial}{\partial z} \left[ w \alpha T^\nu \left( \frac{\partial u_\theta}{\partial z} \right)^\eta \left| \frac{\partial v_\theta}{\partial z} \right|^{\mu-1} \frac{\partial v_\theta}{\partial z} \right], \quad (20)$$

$$\begin{aligned} \frac{\partial T}{\partial t} &= \frac{1}{\rho [c_A (1 - \lambda) + c_B \lambda]} \left[ \alpha T^\nu \left( \frac{\partial u_\theta}{\partial z} \right)^\eta \left| \frac{\partial v_\theta}{\partial z} \right|^{\mu+1} + \frac{k}{w} \frac{\partial}{\partial z} \left( w \frac{\partial T}{\partial z} \right) \right. \\ &\quad \left. + Z \rho [Q + (c_A - c_B) T] (1 - \lambda) \exp \left( -\frac{E}{RT} \right) \right], \end{aligned} \quad (21)$$

$$\frac{\partial u_\theta}{\partial t} = v_\theta, \quad (22)$$

$$\frac{\partial \lambda}{\partial t} = Z (1 - \lambda) \exp \left( -\frac{E}{RT} \right), \quad (23)$$

$$w = w_0 - \frac{h_p}{2} \left[ 1 - \cos \left( \frac{2\pi z}{L_s} \right) \right], \quad (24)$$

where  $Q = e_A^o - e_B^o$  is the heat of reaction.

In order to numerically solve this system of equations, a spatial discretization was performed using second order central differences. The result is a parabolic system of ordinary differential equations in time, as shown in Caspar (1996). The computer code LSODE [Hindmarsh, 1983], the Livermore Solver for Ordinary Differential Equations, was used to step forward in time to solve this system. LSODE solves initial value problems for stiff or nonstiff systems of first order ordinary differential equations using the Gear (1971) method. A stiff system is one whose ratio of largest to smallest eigenvalues in the locally linearized solution matrix is large. That is, the system has rapidly growing or decaying processes that occur over a time scale much shorter than the overall time scale of interest. This computer code is thus desirable for the model presented herein since the process of shear localization occurs over a much shorter time than the overall time of interest. The results of this code were validated in Caspar (1996), through the use of simplified forms of the equations which had exact solutions, and also by comparing results with those of other researchers on previously tested materials.

## **4 Results**

This chapter will present results determined from the TSHB as well as those from the theoretical model. Experimental results for tests on S-7 tool steel (TS) will first be presented, with comparisons drawn between the numerical and experimental results, as well as with results determined from other researchers. Results will then be presented for tests on the following explosive simulants: a PBX cure cast simulant, a PBX pressed simulant, and a melt cast simulant known as Filler-E. These simulants are used to approximate the material properties of PBXN 109, PBX 9501, and tritonal, respectively. The results of these tests will be used to determine approximate parameters for the constitutive law for stress used in this thesis. Finally, numerical simulations will be run on 1018 CRS, and S-7 tool steel to compare with previously determined results, and simulations will be run on the aforementioned explosives.

### **4.1 Experimental Results on Explosive Simulants**

In this section, results determined from the TSHB on tests of the explosive simulants are presented. The strain and strain rate hardening parameters from the constitutive law for stress are then determined from the data. It is important to note that, to our knowledge, no researchers have tested any of these materials in torsion. The results presented in this section are thus previously unrecorded.

#### **4.1.1 Tests on the PBX Cure Cast Simulant**

The results of tests performed at various shear strain rates on the PBX cure cast simulant are included in Figure 8. The reported results are characteristic of a few tests performed at each strain rate. In this figure,

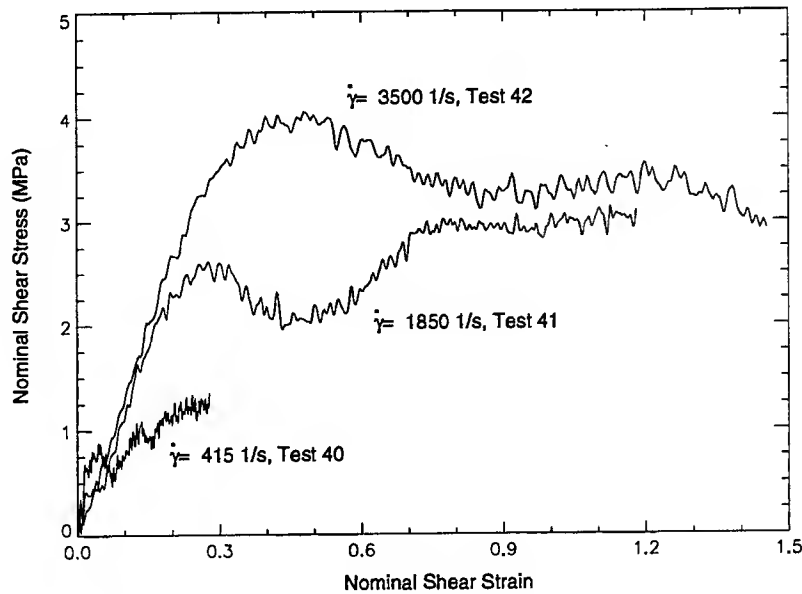


Figure 8: Results from TSHB tests on the PBX cure cast simulant.

test 40 was performed at a shear strain rate of  $415 \text{ s}^{-1}$  and lasted for  $700 \mu\text{s}$ , the full length of the loading pulse. The cause for the initial overshoot in shear stress at a shear strain of 0.05 is unknown, but it could be a material characteristic, or due to the loading geometry. Zener and Hollomon (1944) have stated that a maximum in the stress-strain graph is indicative of the formation of an instability in the deformation. The specimen in this test never failed, and since the shear stress-shear strain curve never reached a maximum, it is assumed that no instability was reached. This test thus provides an accurate measurement of the strain hardening in this material. Prior to performing this test, a line was drawn axially across the specimen. Post test examination of this line revealed no permanent deformation in the specimen. It is thus concluded that this material behaves in a nonlinearly elastic manner over the shear strain rates tested.

Test 41, which was performed at a shear strain rate of  $1850 \text{ s}^{-1}$ , also lasted for  $700 \mu\text{s}$ . The initial peak in the shear stress at a shear strain of 0.28 is a result of the overshoot, as seen in Test 40. Upon post-test examination of the specimen, a small tear was noticed in the circumferential direction within the gage length. This is the consequence of an instability, which could thus account for the decrease in the slope of the shear stress-shear strain curve after a shear strain of 0.75. This test would thus provide an inaccurate measurement of the material's strain hardening characteristic. In addition, it was noticed that even with the onset of instability, the deformation was recovered, indicating purely elastic deformation.

Test 42, which was tested at a shear strain rate of  $3500 \text{ s}^{-1}$ , lasted for about  $400 \mu\text{s}$ . The peak in the shear stress at a shear strain of 0.5 is believed to be a result of the stress overshoot. The specimen in this test failed, with an instability thought to occur around a shear stress of 0.8. This instability prevented the material from further hardening, hence making the overshoot appear to be the occurrence of the instability,

instead of where it is actually thought to occur. Examination of the failure surface revealed voids visible to the naked eye. In addition it was observed that failure did not occur along a single plane, but along an irregular surface, as if the material were torn apart. It is thus doubtful that this material demonstrates shear localization under the given loading conditions.

The data determined herein was then used to calibrate the constitutive law, Equation (9). It is important to note, however, that it is necessary to perform more tests on this material in order to more accurately calibrate the constitutive law. This constitutive law introduces the strain hardening parameter,  $\eta$ , and the strain rate hardening parameter,  $\mu$ . Since it is believed that test 40 results in the most accurate characterization of the strain hardening,  $\eta$  was chosen by trial and error, such that the slope determined from the constitutive model approximately matched the slope of the shear stress-shear strain curve determined from this test. Due to the onset of instability, the results from tests 41 and 42 are unreliable once a maximum shear stress is attained. The results up to the maximum stress are accurate, however, and were used to determine the strain rate hardening parameter,  $\mu$ . These values are tabulated in a later section.

#### 4.1.2 Tests on the PBX Pressed Simulant

Figure 9 shows the results from tests performed by the TSHB on the PBX pressed simulant, where the reported results are characteristic of a few tests performed at each strain rate. Referring to this figure, test

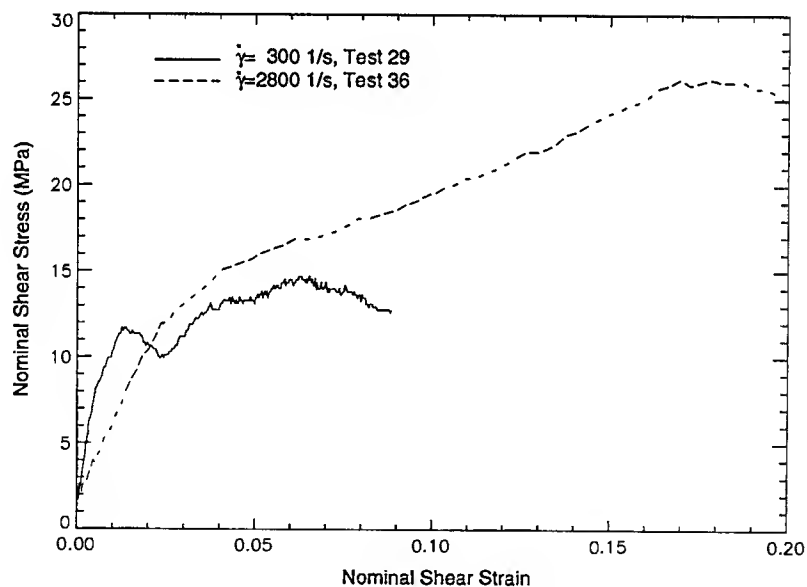


Figure 9: Results from TSHB tests on the PBX pressed simulant.

29, which was performed at a shear strain rate of  $300 \text{ s}^{-1}$ , lasted for about  $350 \mu\text{s}$ . This material in this test also exhibited a stress overshoot, occurring at a shear strain of 0.15. Observation of the post test specimen revealed a planar failure with a rough failure surface including small voids. This is as would be expected in

microvoid nucleation induced shear localization. The instability which caused failure is assumed to account for the peak in the shear stress-shear strain curve at a shear strain of 0.065. Test 38, which was performed at a shear strain rate of  $2800 \text{ s}^{-1}$  lasted only  $75 \mu\text{s}$ , due to the high strain rate deformation. No overshoot was observed in this test. Examination of the post-test specimen revealed fragmentation of the gage length as well as the flanges. Due to this catastrophic failure, these results may not be truly indicative of the material. By a fractographic study of the fragments, it was determined that cracks initiated in the gage length and propagated outward into the flanges at an angle to the cylinder axis, which is indicative of brittle failure.

In order to more accurately determine the order of events in the failure of this specimen, high speed photographs were taken of the deformation experienced by one of these simulants. In the test during which photographs were taken, the shear strain rate was  $2850 \text{ s}^{-1}$ . A plot of the transmitted shear strain for this test, which is proportional to the shear stress in the specimen, is included in Figure 10. High speed

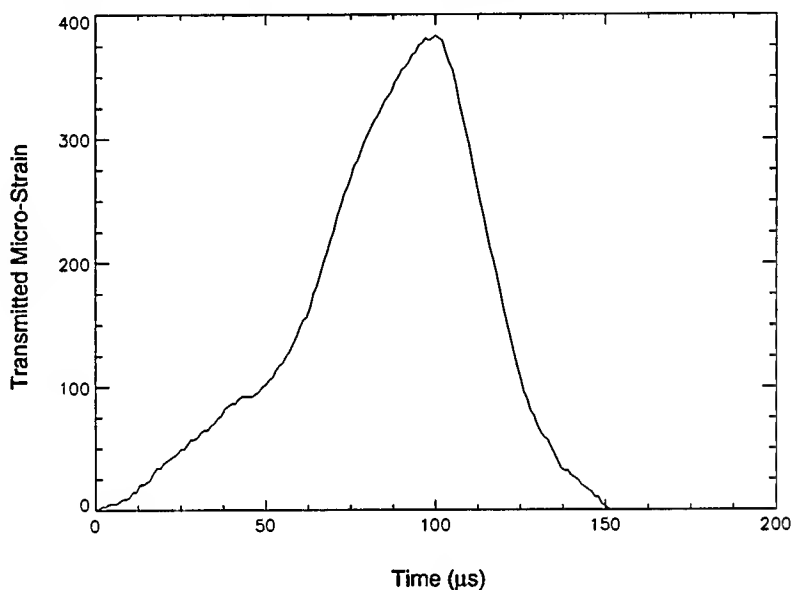


Figure 10: A plot of the transmitted shear strain for Test 49 on the PBX pressed simulant,  $\dot{\gamma} = 2850 \text{ s}^{-1}$ .

photographs of the specimen deformation in this test are included in Figure 11. In these photographs, the vertical black line to the right of the gage length is a result of the camera removing a strip of light from each frame for other purposes. In Figure 11, the photograph labeled  $t = 0 \mu\text{s}$  was taken when the incident pulse first reached the specimen. The photograph labeled  $t = 167 \mu\text{s}$  was taken some time after the transmitted strain, as depicted in Figure 10, reached a maximum. In the center of the gage length of the specimen in this picture, a small crack is visible. From the complete photographic record, not included in this thesis, this crack formed at approximately  $t = 113 \mu\text{s}$ , which is just after the transmitted shear strain reaches a maximum, as seen in Figure 10, which indicates that the drop in the transmitted strain is the result of this failure mechanism. From the photograph labeled  $t = 227 \mu\text{s}$ , it is seen that the crack has increased in size,

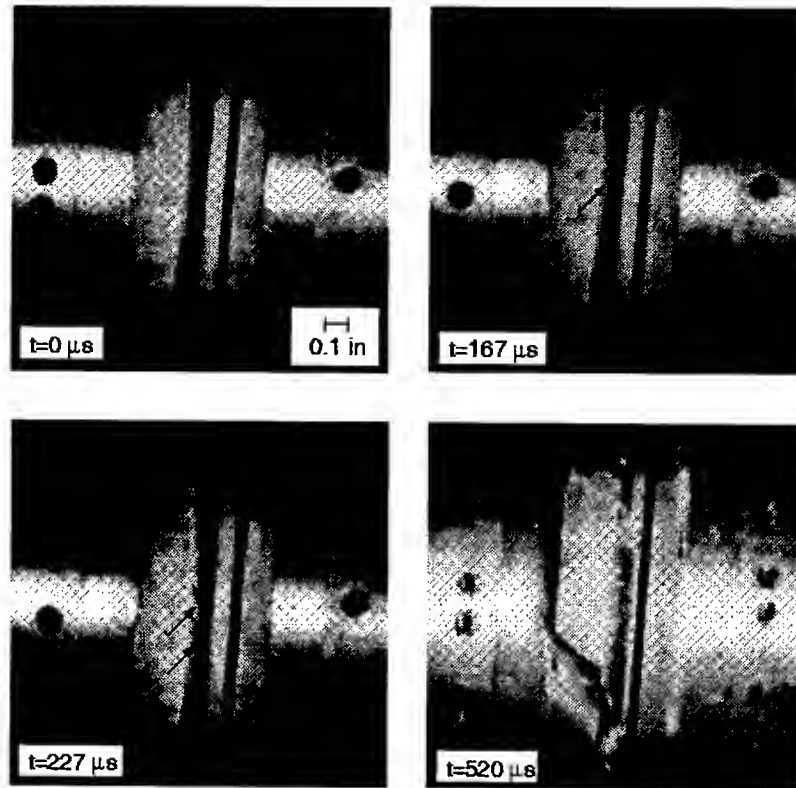


Figure 11: High speed photographs of the failure of a PBX pressed simulant, Test 49.

and that other faint cracks have formed below it. Finally, the photograph labeled  $t = 520 \mu s$ , which was taken of a separate test performed at a comparable shear strain rate, reveals the ultimate fragmentation of the specimen. It is thus confirmed that cracks initiate in the gage length and propagate outward through the flanges. In comparison with the failure of the lower strain rate test, Test 29, it is concluded that the type of failure for this material is dependent on the loading rate.

In order to develop a constitutive model for this material, the parameters from Equation (9) are again approximated to match the constitutive law to this data. The value of  $\eta$  was determined from test 29, since the material in this test exhibited a greater amount of strain hardening prior to the onset of instability. The strain rate hardening parameter,  $\mu$ , was determined such that the constitutive model matched the peak stress obtained in test 38.

#### 4.1.3 Tests on Filler-E

Figure 12 shows the results from tests on Filler-E, where the reported results are characteristics of a few tests performed at each strain rate. In this figure, Test 30, which was performed at a shear strain rate of  $370 s^{-1}$ , lasted  $375 \mu s$ . Fragmentation of the Filler-E specimen occurred in much the same way as did in the high strain rate test on the PBX pressed simulant. Cracks appear to have begun in the gage length



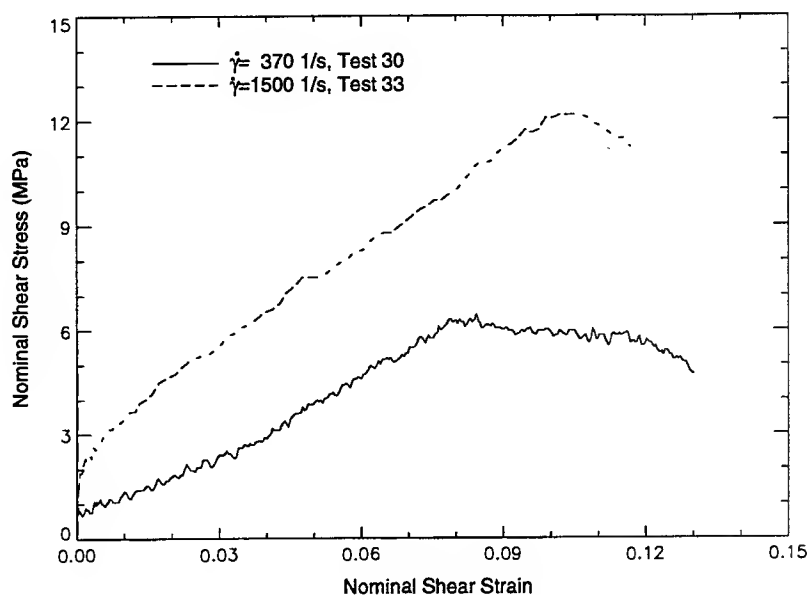


Figure 12: Results from TSHB tests on Filler-E.

and propagated through the specimen at  $45^\circ$  to the axis of the specimen. On tests when there was minimal fragmentation, it was possible to observe the failure surface, which was irregular, not at all indicative of shear localization. On test 33, which was performed at a shear strain rate of  $1500 \text{ s}^{-1}$ , deformation lasted  $120 \mu\text{s}$ . From Figure 12, it is seen that this material has a high sensitivity to strain rate. Tests at this strain rate caused significant fragmentation, with failure similar to that for the low strain rate test. Although Filler-E demonstrates strain and strain rate hardening, the constitutive law could not be fit to a significant portion of the shear stress-shear strain curves seen in Figure 12. This is due to the fact that Filler-E does not demonstrate strain hardening that can be represented in the form of a power law.

## 4.2 Numerical Simulations on Nonreactive Materials

This section presents the results from numerical simulations on PBXN-109 and PBX 9501 with the effects of reaction excluded. The constitutive and material parameters used in these numerical simulations are included in Table 1. The material parameters for PBX 9501 were taken from Dobratz and Crawford (1985), and the material parameters for PBXN-109 and tritonal were taken from Hall and Holden (1988). The thermal conductivity for PBXN-109 was not found, so the value for PBXW-114, a material of similar composition, was used. In order to determine the thermal softening function for some of the materials they tested, Johnson and Cook assumed a linear decrease in the stress as a function of temperature, with the stress reaching zero at the melting point. Since PBX 9501 and PBXN-109 react before melting, their thermal softening parameters are estimated such that the stress is decreased by 50% at the reaction initiation temperature. From Dobratz and Crawford, PBX 9501 reacts at  $240^\circ\text{C}$ , and from Hall and Holden, PBXN-109

Material	Constitutive Parameters				Material Parameters		
	$\alpha$ ( $MPa \frac{s^m}{K^\nu}$ )	$\nu$	$\eta$	$\mu$	$\rho$ ( $\frac{kg}{m^3}$ )	$c_A$ ( $\frac{J}{kg \cdot K}$ )	$k$ ( $\frac{W}{m \cdot K}$ )
PBX 9501	33,000	-1.28	0.320	0.080	1840	1130	0.454
PBXN-109	800	-1.38	0.400	0.320	1670	1260	0.104
Tritonal	-	-9.68	-	-	1690	960	0.460

Table 1: Constitutive and material parameters used in the numerical calculations.

reacts at 220°C. The thermal softening parameter for tritonal was estimated such that the stress is decreased by 90% at the melting temperature of 80°C. The remaining constitutive parameters were determined from the experimental results on the explosive simulants, as described in the previous section.

#### 4.2.1 PBX 9501 Without Reaction

Next, numerical simulations were performed for the deformation of PBX 9501, with the effects of reaction excluded. This was studied in order to compare with the experimental results on the PBX pressed simulant and to determine the material's susceptibility to localization. Reaction was excluded by setting  $Z$  equal to zero in the computer code. The physical constants, included in Table 2 under simulation #1, were chosen to match the experimental conditions. For this case, the test was performed at a shear strain rate of 2800  $s^{-1}$ .

Simulation Number	$v_1$ ( $m/s$ )	$L_s$ ( $mm$ )	$t_1$ ( $\mu s$ )	$w_0$ ( $mm$ )	$h_p$	$v_0$ ( $m/s$ )
1	7.00	2.50	32.14	2.50	0.10	$7.00 \times 10^{-2}$
2	6.25	2.50	32.00	2.50	0.10	$6.25 \times 10^{-2}$

Table 2: Physical constants used in the numerical simulations reported within this thesis.

In order to determine the onset of localization, the following localization criterion, determined by Meyers (1994), is used:

$$\left. \frac{\partial \tau}{\partial \gamma} \right|_{T, \dot{\gamma}} + \left. \frac{\partial \tau}{\partial \dot{\gamma}} \right|_{T, \gamma} \frac{\partial \dot{\gamma} / \partial t|_z}{\partial \gamma / \partial t|_z} \leq \frac{\tau}{\rho c_A} \left. \frac{\partial \tau}{\partial T} \right|_{\gamma, \dot{\gamma}}, \quad (25)$$

which is evaluated at the center of the specimen. The right hand side,  $\Phi$ , and left hand side,  $\Psi$ , of this criterion are plotted as functions of time in Figure 13. In this criterion,  $\Psi$  is a combined measure of the strain and strain rate hardening effects, while the  $\Phi$  is a measure of the thermal softening. It is seen that both  $\Phi$  and  $\Psi$  are always positive, indicating that the material is experiencing strain and strain rate hardening as well as thermal softening, as expected. When  $\Psi$  is less than or equal to  $\Phi$ , this criterion predicts that localization will begin. For this test, the onset of localization is reached after 1.67  $ms$ .

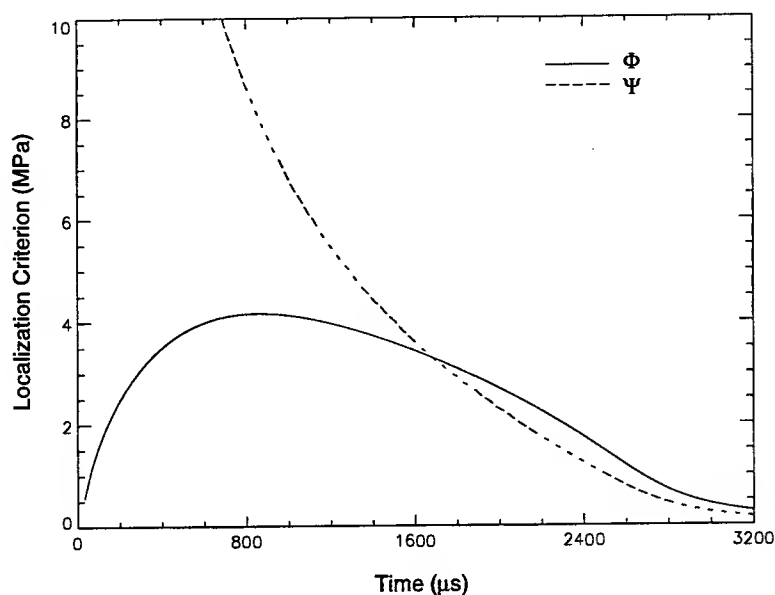


Figure 13: Localization criterion for PBX 9501 without reaction, where  $\Phi$  represents thermal softening and  $\Psi$  represents strain and strain rate hardening.

The effects of localization are readily seen by studying the evolution of the velocity and temperature profile, as seen in Figures 14 and 15. Figure 14 shows the three stage localization process, which was initially observed experimentally by Marchand and Duffy (1988). After the velocity at  $z = L$  reaches its final value, the profile essentially forms a linear distribution in space, which is called homogeneous deformation. Marchand and Duffy have termed this Stage I of the localization process. Since the specimen is thinnest at its center, it is also weakest at that point. The material is thus locally less resistant to deformation, hence developing an inhomogeneous velocity profile with the greatest slope at the center. This is referred to as Stage II. The Stage II localization is very subtle in this test and is not readily observed. Now, the shear strain rate is equal to the slope of the velocity profile, so, as it increases, it causes the stress to increase. The combined effect of the increased shear stress and shear strain rate cause the temperature to increase, as is seen in Figure 15. The rise in temperature causes the stress to drop, which results in further straining and heating. This interaction continues until, after 1.67 ms, the thermal softening dominates over the strain and strain rate hardening. Consequently, deformation rapidly localizes to a narrow region, termed Stage III.

It is this final stage of localization that is termed shear localization, or shear banding. At this time, the rate of change in the temperature increases dramatically at the center of the shear band, resulting in a pronounced spike in temperature. The temperature at the center of the shear band at the onset of localization is 458 K. After 3.2 ms, the temperature at this point has increased to 1590 K. It is interesting to note that the temperature of the material at the onset of localization has already almost reached its initiation temperature of 513 K, and that following localization the temperature far surpasses this value. Hence, it is expected that shear localization in PBX 9501 would produce initiation of reaction.

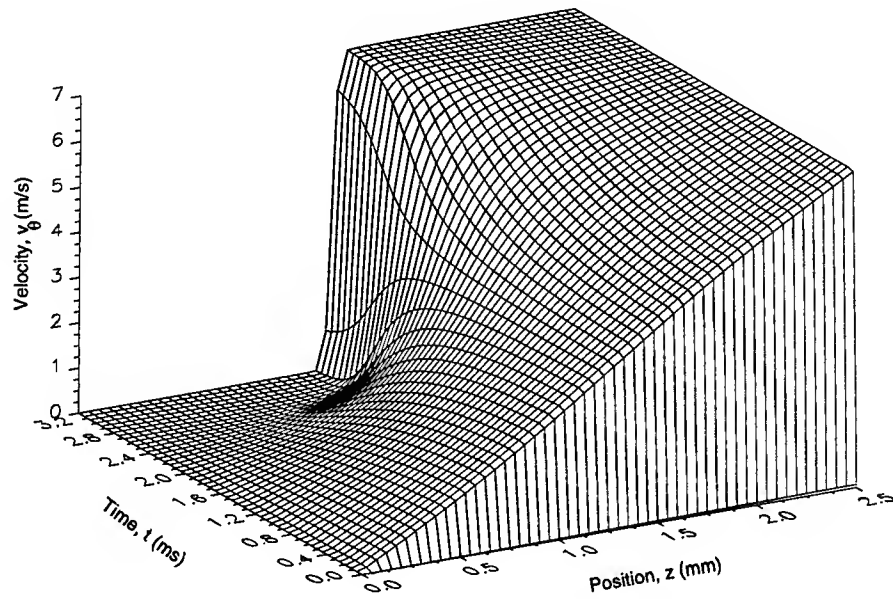


Figure 14: Evolution of the velocity field for PBX 9501 without reaction.

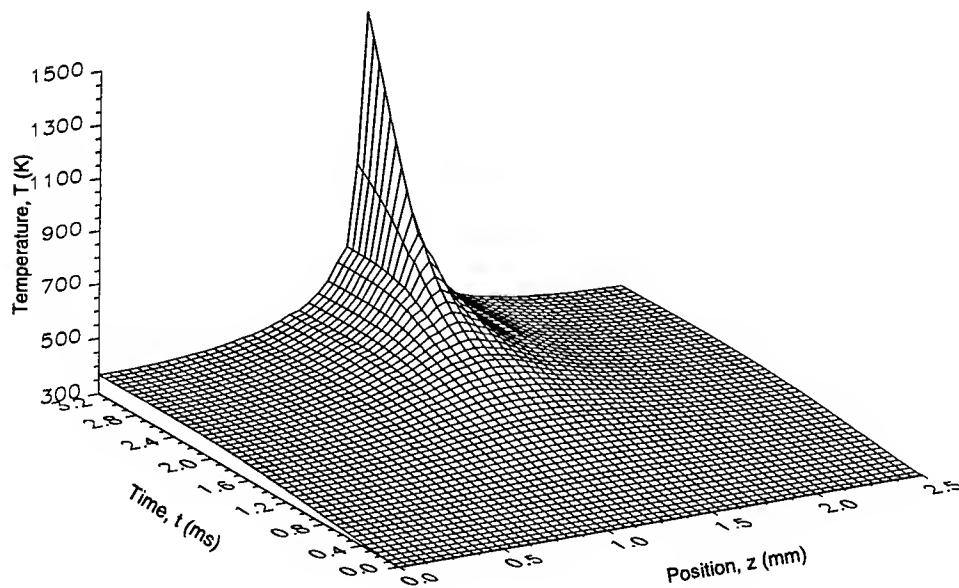


Figure 15: Evolution of the temperature field for PBX 9501 without reaction.

These numerical results are now compared to the experimental results determined from the PBX pressed simulant. Figure 16 compares the experimental and numerical shear stress and shear strain characteristics. From this figure, it is seen that the computer code predicts the shear stress and shear strain characteristics

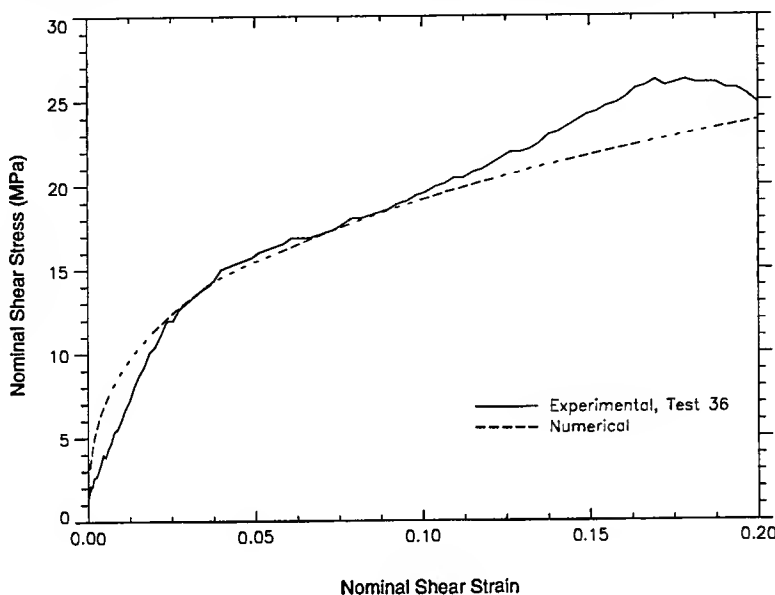


Figure 16: A comparison of the experimental and numerical results for the PBX pressed simulant.

fairly accurately until just before failure. The code, however, does not predict localization to begin until a nominal shear strain of 4.63 is reached, as compared with the experimental failure at 0.20 shear strain. The full numerically determined shear stress-shear strain curve is included in Figure 17. It is thus concluded that the PBX pressed simulant does not fail due to shear localization, but instead due to some other mechanism. This does agree with experimental observations, which suggested that failure could have occurred due to microvoid nucleation and growth, crack propagation or fragmentation. Since these failure mechanisms were not built into the numerical model, the code can not accurately predict this form of failure.

#### 4.2.2 PBXN-109 Without Reaction

Numerical simulations were then performed on PBXN-109, with the effects of reaction excluded. The physical constants, included in Table 2 under simulation #2, were chosen to match the experimental conditions of test 42, with a shear strain rate of  $2500 \text{ s}^{-1}$ . Since it is believed that the PBX cure cast simulant demonstrates nonlinear elastic deformation, it is questionable whether it can be accurately modeled by the numerical method, which assumes viscoplastic heating. Despite this fact, simulations were performed on this material, with the assumption of viscoplastic heating, in order to learn more about the shear localization and thermal initiation processes. The localization criterion for this test is included in Figure 18, which predicts the onset of localization after 272 ms. It is noticed, however, that by the time  $\Phi$  becomes greater than  $\Psi$ ,

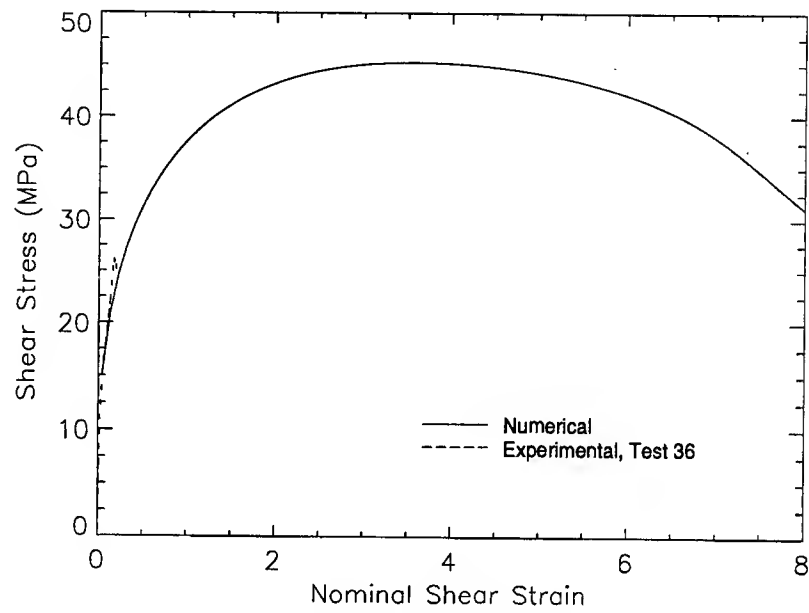


Figure 17: Experimental and numerical shear stress-shear strain curves up to failure for the PBX pressed simulant.

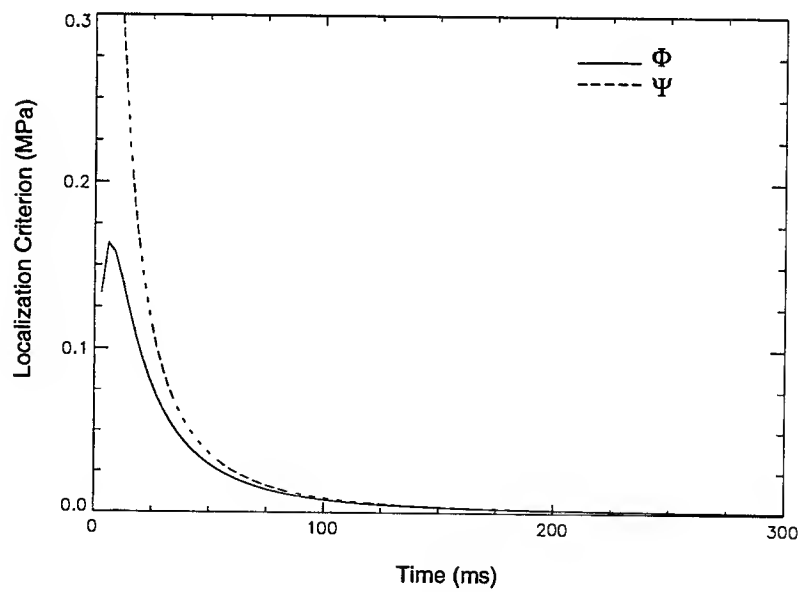


Figure 18: Localization criterion for PBXN-109 without reaction.

the parameters have essentially ceased changing. The shear strain and shear strain rate hardening effects have thus reached a balance with the thermal softening effect. Figure 19 shows the evolution of the velocity

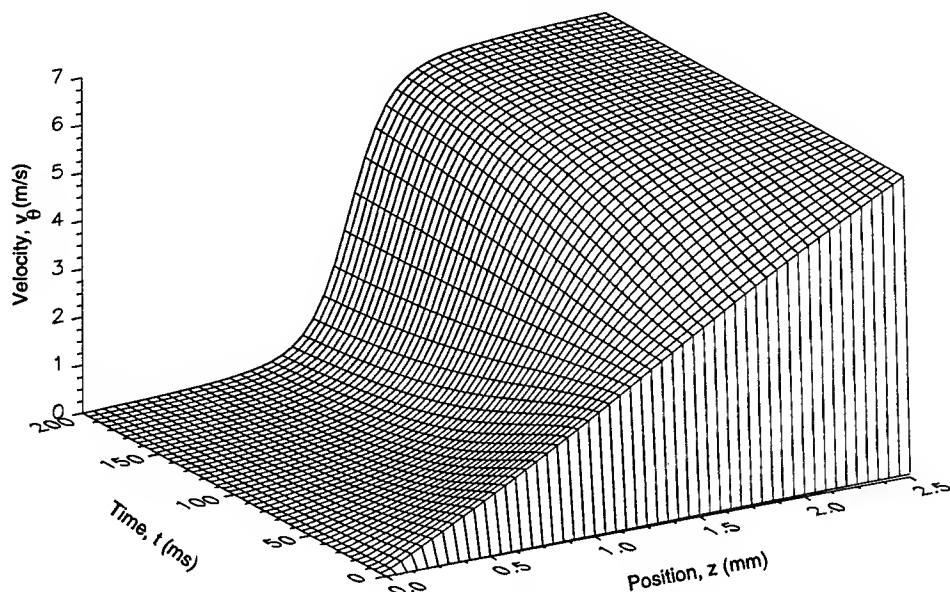


Figure 19: Evolution of the velocity field for PBXN-109 without reaction.

profile for this test. From this figure, it is seen that Stage II is reached, but that transition into localization does not occur. In comparison with PBX 9501, it is seen that PBXN-109 develops a greater inhomogeneity in Stage II, but that PBX 9501 is more susceptible to eventual localization. Further iteration in time reveals that the velocity profile begins to return to a homogeneous state, rather than localizing. Figure 20 shows the evolution of the temperature profile, which shows temperature to increase in an inhomogeneous manner. It is interesting to notice that the temperature exceeds the reaction temperature of  $493\text{ K}$  after about  $10\text{ ms}$ . It is thus concluded that shear localization is not necessary for initiation to occur in this material; an inhomogeneous growth in temperature can eventually lead to significant increases in its value. However, the time over which this process occurs is significantly longer than the experimentally recorded time to failure of  $350\text{ }\mu\text{s}$ . It is hence determined that the PBXN-109 simulant, as did the PBX 9501 simulant, failed experimentally by mechanisms which are not included in this model. Furthermore, it is not expected that deformation in this material can be sustained long enough to increase temperatures into the reactive range.

### 4.3 Numerical Simulations on Reactive Materials

Results are next presented for numerical simulations on reactive materials. The reactive parameters for PBX 9501, PBXN-109 and tritonal are included in Table 3. Dobratz and Crawford (1985) list maximum calculated and experimentally determined values of  $Q$  for various explosives. They also tabulate values of  $Z$  and  $E$  for various explosives, but not for any of those tested herein. For PBX 9501, Dobratz and

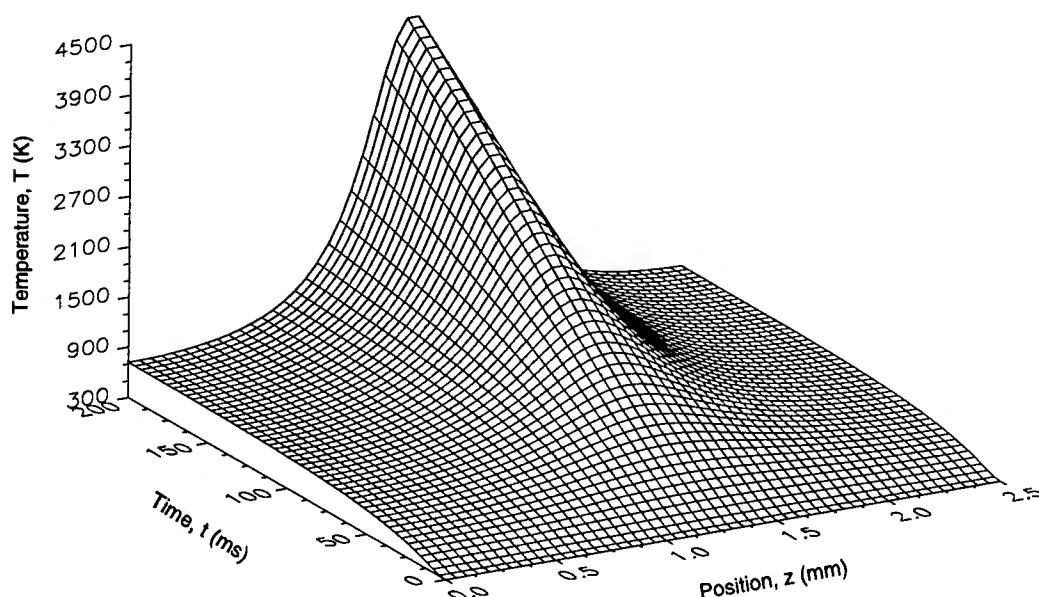


Figure 20: Evolution of the temperature field for PBXN-109 without reaction.

Crawford only list the maximum calculated value of  $Q$ . For PBX 9404, a similar explosive, the experimental value is 88.5% of the maximum, hence, the  $Q$  value for PBX 9501 is chosen to be 88.5% of its maximum calculated value. Since PBX 9501 is 95% by weight HMX, the values of  $Z$  and  $E$  for HMX were used. Neither PBXN-109 nor tritonal are included in Dobratz and Crawford's handbook. Since PBXN-109 is 64% RDX, the experimentally determined value of  $Q$  and the values of  $A$  and  $E$  for RDX were used to simulate this explosive. Likewise, as tritonal is 80% TNT, the values of  $Q$ ,  $Z$ , and  $E$  for TNT were used to simulate tritonal.

Material	$Z$ ( $s^{-1}$ )	$Q$ ( $kJ/kg$ )	$E$ ( $kJ/mol$ )	$R$ ( $J/mol \cdot K$ )
PBX 9501	$5.00 \times 10^{19}$	5891	220.6	8.314
PBXN-109	$2.02 \times 10^{18}$	6320	197.1	8.314
Tritonal	$2.51 \times 10^{11}$	4560	143.9	8.314

Table 3: Reactive constants used in the numerical code.

#### 4.3.1 PBX 9501 With Reaction

Results are now presented for simulations of PBXN-109 with the effects of reaction included. The same parameters were used in this simulation as in the nonreactive case. The effects of including reaction proved to have little effect on the results prior to initiation. The localization criterion predicted localization after 1.682 ms, a difference of only 0.3% from the nonreactive case. The temperature at the center of the



specimen at this time was within 0.1% of the corresponding nonreactive temperature. As was anticipated by the nonreactive case, reaction in the reactive test did occur shortly following the onset of localization. The evolution of the velocity and temperature profiles for this material appeared very similar to those of the nonreactive case. Computation was stopped in this simulation shortly following the start of reaction, since the reaction proceeded so quickly that the time step rapidly approached zero.

The temperature profile for this simulation is similar to that of the PBX pressed simulant, with reaction occurring prior to severe development of the temperature spike. By observing the evolution of the reaction progress variable, Figure 21, it is seen how sensitive initiation is to temperature. Appreciable reaction did not

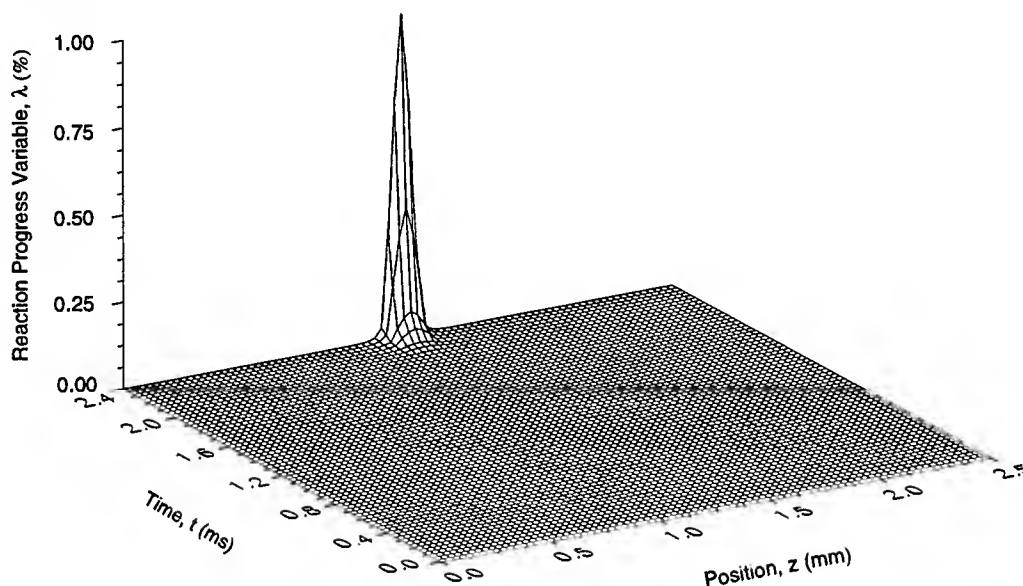


Figure 21: Evolution of the reaction progress for PBX 9501 with reaction.

begin until the reaction temperature was reached, at which time reaction quickly initiates in the localized hot spot. Bowden and Yoffe (1985) have discussed initiation in the context of localized hot spots, shear localization being only one of the mechanisms by which hot spots are generated. Also, it is observed how quickly reaction proceeds, with the reaction increasing over 10 times its value in the last 25  $\mu s$ , to achieve approximately 1.0% completion at the center of the specimen. It is important, however, to state that the nominal shear strain reached at initiation is approximately 6.4, whereas the simulant failed after a shear strain of 0.2 experimentally. As stated previously, experimental observations suggested failure by other mechanisms, which are not included in this numerical code.

#### 4.3.2 PBXN-109 With Reaction

The final material studied in this thesis is PBXN-109 with reaction effects included. Again, the same parameters were used in this simulation as in the nonreactive case. A study of the velocity profile up to

reaction initiation reveals that no significant inhomogeneity has yet to develop. In contrast to the velocity profile, the temperature profile develops an inhomogeneity. In fact, the temperatures in this inhomogeneity, although not significant, have grown enough to reach the reaction temperature. The plot of the reaction progress variable profile (Figure 22) reveals, as expected from the nonreactive case, that reaction has occurred

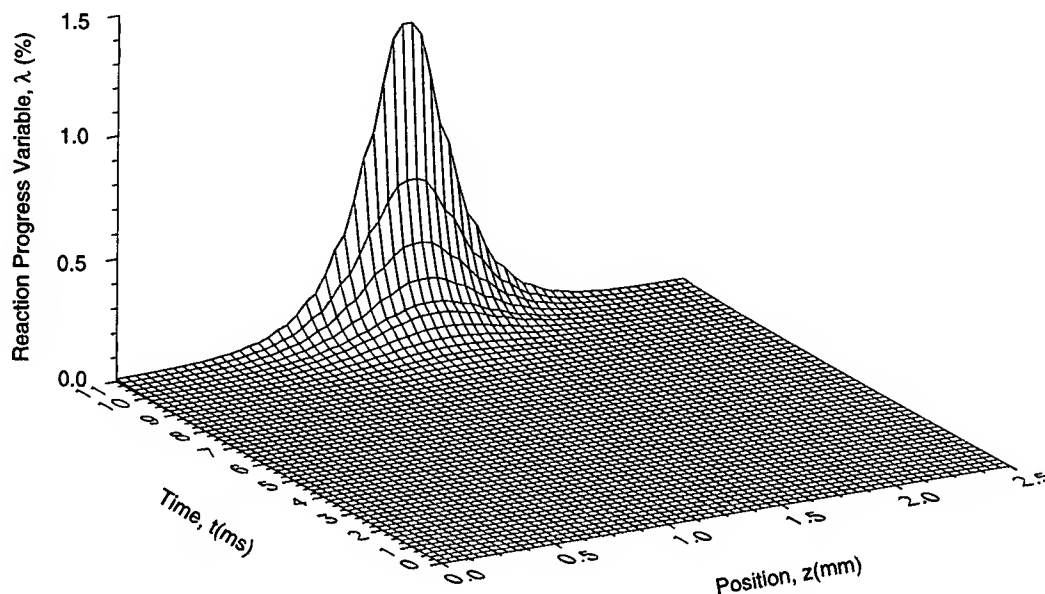


Figure 22: Evolution of the reaction progress variable profile for PBXN-109 with reaction.

prior to the onset of localization. In comparison with the results from PBX 9501, however, it is seen that reaction occurs over a much larger spatial region. This is due to the fact that the temperature is not severely localized when it reaches the initiation temperature. Computations were stopped after reaction reached approximately 1.3% completion, due to the explosive growth in reaction and consequent decrease in time step to zero. It is thus concluded that shear localization is not necessary to produce initiation. A mere inhomogeneity, if allowed to grow long enough, can result in the initiation of reaction. For this to occur, however, all other forms of failure would have to be suppressed. The nominal shear strain reached at initiation in this simulation, was 26.7, far in excess of the experimentally determined nominal shear strain of 1.5, which was reached at failure. As was the case in the simulation on PBX 9501, it is determined that the current analytical method is insufficient for modeling failure in PBXN-109.

## 5 Conclusions

The results included herein first present constitutive behavior in the form of shear stress-shear strain curves for a PBX cure cast simulant, a PBX pressed simulant, and a melt cast simulant known as Filler-E. These simulants are used to approximate the material properties of PBXN 109, PBX 9501, and tritonal,

respectively. Results from these tests revealed significant dependencies of the shear stress on shear strain and shear strain rate, as compared with the corresponding dependencies of steels. Observation of the failure surface of the various explosive simulants revealed evidence of tearing, microvoid nucleation and growth, crack propagation and fragmentation. The data from these tests can be used to calibrate various constitutive laws which are used to input experimental data into analytical codes. The data from the PBX cure cast and pressed simulants has been used to calibrate a constitutive law, proposed by Clifton, *et al.* (1984), which models shear stress by including the effects of strain and strain rate hardening as well as thermal softening.

From numerical simulations on PBX 9501 without the effects of reaction, the three stage localization process observed experimentally by Marchand and Duffy (1988) was predicted, with the onset of localization predicted after a nominal shear strain of 4.63. In addition, the subsequent rise in temperature quickly exceeding the reaction temperature of the material. When the effects of reaction were included, initiation of reaction began after a nominal shear strain of 6.4. From simulations on PBXN-109 without the effects of reaction, it was determined that this material is not susceptible to localization. With the inclusion of reaction effects, however, it was determined that reaction could occur without localization. Due to a mere growth in the inhomogeneous temperature field, reaction initiating after a nominal shear strain of 26.7.

Since localization is assumed to be followed by failure or initiation, numerical results agreed with experimental results in predicting that PBX 9501 would fail at a lower strain than PBXN-109. Experimentally, the corresponding simulants failed after nominal shear strains of 0.2 and 1.5, respectively. In comparison of the experimental and numerical values, however, it is seen that the experimental tests failed at significantly lower shear strains. This is not surprising, since experimental observations indicated that failure could have occurred as a result of the combined mechanisms of microvoid nucleation and growth, crack propagation and fragmentation. These mechanisms are not included in the current analytical study and hence it is not possible for their results to be predicted.

It is, however, concluded that if these other mechanisms were suppressed, localization and/or initiation would occur in the tested materials. Chou *et al.* (1991) stated that brittle materials become more ductile under the application of hydrostatic stresses. In addition, Frey (1981) concluded that explosives under compressive stresses generate more heat when being deformed, hence decreasing the time necessary for initiation to occur. Furthermore, Chou *et al.* concluded that localization becomes significant in covered explosives. Finally, Dodd and Atkins (1983) concluded that increased hydrostatic stresses tended to decrease microvoid nucleation. The explosives in deep earth penetrators are contained and subject to an unknown amount of hydrostatic stress. Deformation under these circumstances could thus result in the suppression of failure mechanisms and hence increase the susceptibility to localization and initiation. As a result, it is desired to perform future tests under the application of hydrostatic stresses, in order to determine the subsequent effect on localization and initiation.

It is also concluded in Caspar (1996) that the most important parameters in the study of shear localization for a given material are the constitutive parameters. In order to increase the mass, and hence momentum, of reactive devices, studies are being performed at Eglin Air Force Base on an explosive unofficially termed TUNG-5. This explosive uses tungsten as a binder for the explosive crystals. Due to its high strength, it is concluded in this thesis that such a material would be particularly susceptible to shear localization and hence reaction. This material would have the material characteristics of a metal and the reactive characteristics of an explosive. Since it is known that metals are particularly susceptible to localization, and since this thesis has shown that localization is quickly proceeded by initiation, significant precautions should be taken in the development of munitions containing this material.

For future work, several ideas are presented which may obtain more accurate results from the TSHB. First, the incident and transmission bars will be ground straight. This will reduce any bending and axial pulses, as well as decrease inhomogeneous deformation in the specimen. It is also desired to develop a better method for aligning the bars. Currently, delrin bearings are used to support the bars. With harder bearings, the bars can be restrained from bending motion; in addition, it will be easier to determine proper alignment of the bars by observing the amount of resistance when the bars are rotated by hand. Finally, since preliminary results with new strain gages revealed up to 8% difference in transmitted strain, it is desired to apply new strain gages to the bars. Finally, it is desired to calibrate the amplifiers in order to determine a measure of their accuracy.

## References

- [1] R. C. Batra and C. H. Kim (1991), "The Effect of Thermal Conductivity on the Initiation, Growth, and Band Width of Adiabatic Shear Bands," *Int. J. Eng. Science*, Vol. 29, pp. 949-960.
- [2] R. C. Batra and C. H. Kim (1992), "Analysis of Shear Banding in Twelve Materials," *International Journal of Plasticity*, Vol. 8, pp. 425-452.
- [3] R. C. Batra, X. Zhang, and T. W. Wright (1995), "Critical Strain Ranking of 12 Materials in Deformations Involving Adiabatic Shear Bands," *Journal of Applied Mechanics*, Vol. 62, pp. 252-255.
- [4] F. P. Bowden and Y. D. Yoffe (1985), *Initiation and Growth of Explosives in Liquids and Solids*, Cambridge University Press, Cambridge, Great Britain.
- [5] V. Boyle, R. Frey, and O. Blake (1989), "Combined Pressure Shear Ignition of Explosives," *Ninth Symposium (International) on Detonation*, pp. 3-17.
- [6] R. J. Caspar (1996), "Experimental and Numerical Study of Shear Localization as an Initiation Mechanism in Energetic Solids," Masters Thesis, University of Notre Dame.
- [7] P. C. Chou, W. Flis, and D. Jann (1991), "Explosive Response to Unplanned Stimuli," Dyna East Corporation Technical Report DE-TR-91-15.
- [8] I. G. Currie (1993), *Fundamental Mechanics of Fluids*, 2nd ed., McGraw-Hill, Inc., New York, pp. 224-228.

- [9] R. J. Clifton, J. Duffy, K. A. Hartley, and T. G. Shawki (1984), "On Critical Conditions for Shear Band Formation at High Strain Rates," *Scripta Met.*, Vol. 18, pp. 443-448.
- [10] L. S. Costin, E. E. Crisman, R. H. Hawley, and J. Duffy (1979), *2nd Conference on the Mechanical Properties of Materials at High Rates of Strain*, Ed. by J. Harding, The Institute of Physics, London, 90.
- [11] B. M. Dobratz and P. C. Crawford (1985), *LLNL Explosives Handbook-Properties of Chemical Explosives and Explosive Simulants*, Lawrence Livermore National Labs., UCRL-52997, National Technical Information Service, DE91-006884.
- [12] B. Dodd and A. G. Atkins (1983), "Flow Localization in Shear Deformation of Void-Containing and Void-Free Solids," *Acta Metall.*, Vol. 31, pp 9-15.
- [13] J. Duffy and Y. C. Chi (1992), "On the measurement of local strain and temperature during the formation of adiabatic shear bands," *Materials Science and Engineering*, A157, pp. 195-210.
- [14] W. Fickett and W. C. Davis (1979), *Detonation*, University of California Press, Berkeley, CA.
- [15] J. E. Field, G. M. Swallowe and S. N. Heavens (1982), "Ignition Mechanisms of Explosives during Mechanical Deformation," *Proc. R. Soc. Lond. A* 382, pp. 231-244.
- [16] R. B. Frey (1981), "The Initiation of Explosive Charges by Rapid Shear," *Seventh Symposium (International) on Detonation*, Naval Surface Weapons Center, Annapolis, MD, pp. 36-42.
- [17] R. B. Frey (1985), in *Eighth Symposium (International) on Detonation*, Naval Surface Weapons Center, Albuquerque, NM, pp. 68-80.
- [18] J. H. Giovanola (1988, a), "Adiabatic Shear Banding Under Pure Shear Loading. Part I: Direct Observation of Strain Localization and Energy Dissipation Measurements," *Mechanics of Materials*, Vol. 7, pp. 59-71.
- [19] J. H. Giovanola (1988, b), "Adiabatic Shear Banding Under Pure Shear Loading. Part II: Fractographic and Metallographic Observations," *Mechanics of Materials*, Vol. 7, pp. 73-87.
- [20] T. N. Hall and J. R. Holden (1988), *Navy Explosives Handbook. Explosion Effects and Properties-Part III. Properties of Explosives and Explosive Compositions.*, Naval Surface Warfare Center, NSWC MP 88-116, Defense Technical Information Center, AD-B138 762.
- [21] J. Harding, E. D. Wood, and J. D. Campbell (1960), "Tensile Testing of Material at Impact Rate of Strain," *J. Mech. Eng. Sci.*, Vol. 2, p. 88.
- [22] K. A. Hartley and J. Duffy (1985), "Introduction, High Strain Rate Shear Testing" in "High Strain Rate Testing, Mechanical Testing" *Metal Hand Book*, American Society for Metals, Vol. 8, Edition 9, p. 215.
- [23] K. A. Hartley, J. Duffy, and R. H. Hawley (1985), "The Torsional Kolsky (Split-Hopkinson) Bar, High Strain Rate Shear Testing" in "High Strain Rate Testing, Mechanical Testing," *Metal Hand Book*, American Society for Metals, Vol. 8, Edition 9, pp. 218-230.
- [24] K. A. Hartley, J. Duffy, and R. H. Hawley (1987), "Measurement of the Temperature Profile During Shear Band Formation in Steels Deforming at High Strain Rates," *J. Mech. Phys. Solids*, Vol. 35, No. 3, pp. 283-301.
- [25] A. C. Hindmarsh (1983), "ODEPACK, A Systematized Collection of ODE Solvers," *Scientific Computing*, R. S. Stepleman *et al.*, Eds., IMACS/North-Holland Publishing Company, Amsterdam, pp. 55-64.

- [26] G. R. Johnson and W. H. Cook (1983), "A Constitutive Model and Data for Metals Subjected to Large Strains, High Strain Rates and High Temperatures," *Proc. 7th Int. Symp. Ballistics*, The Hague, The Netherlands, pp. 541-548.
- [27] J. Kang, P. B. Butler, and M. R. Baer (1992), "A Thermomechanical Analysis of Hot Spot Formation in Condensed-Phase, Energetic Materials," *Combustion and Flame*, Vol. 89, pp. 117-139.
- [28] H. Kolsky (1949), "An Investigation of the Mechanical Properties of Materials at Very High Rates of Loading," *Proc. Phys. Soc. London*, Vol. 62-B, pp. 676-700.
- [29] H. Kolsky (1953), *Stress Waves in Solids*, Oxford University Press, London.
- [30] J. Lubliner (1990), *Plasticity Theory*, Macmillan Publishing Co., New York, pp. 69-99.
- [31] A. Marchand and J. Duffy (1988), "An Experimental Study of the Formation Process of Adiabatic Shear Bands in a Structural Steel," *J. Mech. Phys. Sol.*, Vol. 36, No. 3, pp. 251-283.
- [32] K. G. McConnell and W. F. Riley (1993), "Strain-Gage Instrumentation and Data Analysis," in *Handbook on Experimental Mechanics*, A. S. Kobayashi, Ed., 2nd edition, VCH Publishers, Inc., New York, NY, pp. 79-117.
- [33] L. W. Meyer (1992), "Constitutive Equations at High Strain Rates," in *Shock-Wave and High-Strain-Rate Phenomena in Materials*, M. A. Meyers, L. E. Murr, and K. P. Staudhammer, Eds., Marcel Dekker, Inc., New York, NY, pp. 49-68.
- [34] M. A. Meyers (1994), *Dynamic Behavior of Materials*, John Wiley & Sons, Inc., New York, NY.
- [35] M. A. Meyers, L. E. Murr, K. P. Staudhammer (1992), *Shock-Wave and High-Strain-Rate Phenomena in Materials*, Marcel Dekker, Inc., New York, NY.
- [36] A. M. Rajendran (1992), "High Strain Rate Behavior of Metals, Ceramics and Concrete," Report # WL-TR-92-4006, Wright Patterson Air Force Base.
- [37] H. C. Rogers (1979), "Adiabatic Plastic Deformation," *Ann. Rev. Mater. Sci.*, Vol. 9, pp. 283-311.
- [38] T. Weerasooriya (1990), "The MTL Torsional Split-Hopkinson Bar," U. S. Army Materials Technology Laboratory Report MTL TR 90-27.
- [39] G. B. Whitham (1974), *Linear and Nonlinear Waves*, John Wiley & Sons, Inc., New York, NY.
- [40] T. W. Wright (1987), "Steady Shearing in a Viscoplastic Solid," *J. Mech. Phys. Solids*, Vol. 35, No. 3, pp. 269-282.
- [41] T. W. Wright and R. C. Batra (1985), "Further Results on the Initiation and Growth of Adiabatic Shear Bands at High Strain Rates," *Journal de Physique*, Coll. C5, suppl. to no. 8, p. C5-323.
- [42] C. Zener and J. F. Hollomon (1944), "Effect of Strain Rate upon Plastic Flow of Steel," *Journal of Applied Physics*, Vol. 15, pp. 22-32.

A Molecular-Level View of Solvation in Supercritical Fluid Systems

Emily D. Niemeyer  
Graduate Research Assistant  
Department Chemistry

State University of New York at Buffalo  
Buffalo, NY

Final Report for:  
Summer Research Extension Program  
Wright Laboratory

Sponsored by:  
Air Force Office of Scientific Research  
Bolling AFB, Washington, DC

and

Wright Laboratory

December 1996

## A MOLECULAR-LEVEL VIEW OF SOLVATION IN SUPERCRITICAL FLUID SYSTEMS

Emily D. Niemeyer  
Graduate Research Assistant  
Department of Chemistry  
State University of New York at Buffalo

### Abstract

Although supercritical fluids (SFs) have generated intense research interest within recent years, there is still much to be learned about the intermolecular interactions which occur between a SF and a dissolved solute. We have conducted a series of experiments in both neat and modified supercritical fluid systems to determine the fundamental interactions that occur in the local environment surrounding a solute molecule. Specifically, we have used steady-state and time-resolved fluorescence spectroscopy as a tool to provide information on solute-fluid interactions in supercritical water, supercritical alkanes, supercritical CO<sub>2</sub>-diluted polymers, and cosolvent-modified supercritical CO<sub>2</sub>. Solvation phenomena surrounding an organic solute (pyrene) in near and supercritical water have been quantified using combined steady-state and time-resolved fluorescence measurements. Pyrene has also been used to determine molecular-level interactions occurring surrounding a solute dissolved in supercritical *n*-alkanes (supercritical fuel precursors). Rotational reorientation measurements have been used to quantify the effect of a supercritical fluid on solute dynamics within a model polymer system. Rotational reorientation measurements have also been used to determine the extent and magnitude of solute-fluid clustering occurring in cosolvent-modified CO<sub>2</sub>.



## A MOLECULAR-LEVEL VIEW OF SOLVATION IN SUPERCRITICAL FLUID SYSTEMS

Emily D. Niemeyer

### Introduction

The ability to tune the physicochemical properties of supercritical fluids (e.g., refractive index ( $n$ ), density ( $\rho$ ), and dielectric constant ( $\epsilon$ )) offers a great advantage for many applications ranging from chromatography and extractions<sup>1-5</sup> to processing and synthesis.<sup>6-13</sup> Below the fluid critical point (defined by its characteristic critical temperature and pressure), the liquid and gas phases can exist in equilibrium. However, above the critical point, the two phases coalesce into a single phase that exhibits many of the desirable properties of both gases and liquids - a supercritical fluid. For example, supercritical fluids possess favorable mass transport and increased solvation, and one can adjust these parameters by very slight changes in temperature and pressure.<sup>14</sup> Thus, supercritical fluids are often thought of as completely tunable solvents.

Supercritical fluids have been used industrially in a variety of fields ranging from hazardous waste disposal<sup>3,15-20</sup> to polymer processing.<sup>2</sup> Because many supercritical fluids are environmentally benign (e.g., CO<sub>2</sub> and H<sub>2</sub>O), they have also become the focus of intensive research aimed at replacing hazardous and/or expensive organic solvents currently in use. In the food and beverage industry, supercritical CO<sub>2</sub> has replaced organic solvents for everything from decaffeination of coffee to the removal of flavor essences from spices.<sup>23</sup> Supercritical CO<sub>2</sub> has also been used by the petroleum industry for removal of contaminants from coal and the synthesis and processing of commonly used fluoropolymers.<sup>2</sup> More recently, significant interest has focused on supercritical water oxidation as a methodology for the disposal of hazardous waste and chemical arms.<sup>15-20</sup>

It has been well-established both theoretically<sup>21,22</sup> and experimentally<sup>23-26</sup> that an increase in fluid density surrounding a solute occurs in the proximity of the fluid critical point. This increased interaction is often termed "solute-solvent clustering"<sup>23</sup> or "molecular charisma".<sup>24</sup> These solute-fluid clusters are dynamic in nature, constantly exchanging fluid molecules with the bulk on a short time scale.<sup>21</sup> In addition, the maxima in clustering has recently been determined to occur at near one half of the critical density, much lower than previously thought.<sup>22</sup> This clustering phenomena is known to affect reaction rates and outcome,<sup>6,7</sup> solute conformational equilibria,<sup>26</sup> and extraction processes.<sup>5</sup> Therefore, in order to fully exploit the potential of supercritical fluids, one must develop a molecular-level view of solvation in supercritical solvents like CO<sub>2</sub> and H<sub>2</sub>O.

A portion of the Air Force mission has aimed toward the development of newer, high-speed aircraft. These advanced aircraft will rely heavily on the onboard fuel supply as the major means to cool the plane fuselage. As a result, at any given time, portions of the fuel may be raised above its critical temperature. Therefore, questions about the fuel stability, the internal fuel dynamics and interactions with dissolved matrix concomitants (additives), and the exchange properties *under supercritical conditions* represent key factors that will govern the ultimate performance of the fuels

and, hence, these advanced, high-speed aircraft. For example, if enhanced solute-fluid, solute-cosolvent, and solute-solute interactions occur and persist in or near the fuel critical point, it is entirely possible for the performance of a given fuel to plummet and even for the fuel delivery system to fail entirely. For this reason, it is imperative to develop a more comprehensive molecular-level understanding of the nature of the interactions occurring within supercritical fluid and supercritical fuel systems.

Toward this end, we have conducted several studies to determine solute-fluid interactions occurring in neat and modified supercritical fluid systems. Steady-state and time-resolved fluorescence measurements have been used to experimentally quantify solvation phenomena occurring in near and supercritical water using the fluorescent probe pyrene (*Appl. Spectrosc.*, submitted for publication). We have also determined the extent of solute fluid clustering for pyrene dissolved in supercritical *n*-alkanes (fuel precursors). Rotational reorientation measurements have been used to quantify the effect of a supercritical fluid on solute dynamics within a model polymer system (manuscript in preparation). Rotational reorientation measurements have also been used to determine the extent and magnitude of solute-fluid clustering in cosolvent-modified CO<sub>2</sub>. The remainder of this document expands on each of these research areas.

### **Probing Molecular-Level Interactions Occurring in SCW**

Supercritical water (SCW) has received much attention in recent years because it serves a key roll in hazardous waste disposal and chemical arms destruction.<sup>15-20</sup> For example, environmentally harmful organics can be rapidly and efficiently oxidized in SCW to benign compounds such as CO<sub>2</sub>, H<sub>2</sub>O, and simple inorganic salts and acids.<sup>16-19,27</sup> Recently, the first commercial hazardous waste disposal facility based on SCW went on line in Austin, TX, proving that SCW oxidation is suitable for large scale processing.<sup>17,19</sup> Supercritical water oxidation plants have also been constructed, as an alternative to incineration of municipal sludge and industrial wastes, in Germany, Canada and Japan.<sup>16</sup>

Although water is likely the most well-studied chemical solvent, when it is raised above its critical point ( $T_c = 374.4$  °C,  $P_c = 220.55$  bar, and  $\rho_c = 0.281$  g/mL), it becomes a solvent with fascinating properties.<sup>27,30,31</sup> Specifically, liquid water is clearly a polar solvent that dissolves well ionic species and dissolves less well hydrophobic solutes. SCW, in contrast, behaves more like a “nonaqueous” solvent, becoming completely miscible with nonpolar compounds that are relatively insoluble in liquid water at ambient conditions.<sup>27</sup> Like other supercritical fluids, SCW also has physicochemical properties that can be tuned between gas and liquid-like values, making it an attractive medium for new reactions and waste disposal.<sup>30,31</sup> Water above its critical point can also become extremely corrosive and even dissolves common materials such as stainless steel and quartz.<sup>27,32</sup>

Although recent interest in SCW processing has soared, research on the fundamental interactions that occur in SCW is lagging. For example, there have been many studies aimed toward understanding the kinetics and mechanism of basic reactions in SCW,<sup>28-29,33-39</sup> but there have been relatively few experiments aimed at quantifying the molecular-level interactions between the solute and the fluid. Molecular dynamics calculations<sup>40</sup> have been used to model SCW and supercritical aqueous solutions. These results show that SCW can maintain a solvation shell similar to that seen in

ambient liquid water, with long range solvation phenomena similar to those observed in other supercritical fluids. Interestingly, although solute-fluid clustering has been reported using molecular simulations in SCW,<sup>40</sup> the extent of density augmentation surrounding an organic solute has not been fully quantified experimentally. Also, although solute-fluid clustering is common in other fluid systems,<sup>21-26</sup> recent molecular dynamics work by Gao<sup>41</sup> suggests solute-fluid rarefaction (i.e., a decrease in the local density of solvent molecules surrounding the solute) for the benzene dimer in SCW.

We have aimed to quantify experimentally the local density surrounding a model organic solute dissolved in SCW. Toward this end, we use steady-state and time-resolved fluorescence to provide a molecular-level view of the nature of solvation occurring in SCW and to compare these interactions with those observed in other supercritical fluids.

Many of the reasons for the lack of spectroscopic data in SCW are associated with the system's high temperature and pressure. Pyrene fluorescence has been studied in water to near-critical temperatures and these reports state that pyrene is stable for up to 2 hrs at 345 bar.<sup>42</sup> For this reason and others (*vide infra*), we have chosen pyrene as our model organic solute. Fluorescence from pyrene has been used to probe local environments in a variety of media,<sup>43-45</sup> including supercritical fluids<sup>42,46-50</sup> and its photophysics are well known. For example, the intensity of the 0-0 transition ( $I_1$ ) is solvent dependent while the 0-3 transition ( $I_3$ ) is solvent insensitive. Thus, the  $I_1/I_3$  ratio provides a means to quantify the local environment surrounding the pyrene molecule.<sup>43</sup> This approach has been used previously in supercritical CO<sub>2</sub> to determine the degree of local density augmentation over a broad density region.<sup>46,47</sup> However, because of the nature of SCW, these measurements become significantly more challenging because of temperature-induced broadening of the pyrene emission spectrum. Finally, one must carefully deoxygenate fully the pyrene/water solutions to minimize solute oxidation.

## Experimental

### Reagents and Sample Preparation

Pyrene (99.9%) was purchased from Aldrich and used as received. Deionized (18 M $\Omega$ ) ultra filtered water was used without further purification. Samples of pyrene in water were prepared by stirring a saturated solution of pyrene for several days and filtering prior to use with a fritted funnel. This results in a solution with a concentration of pyrene which is approximately 0.5  $\mu$ M.<sup>51</sup> There were not indications of aggregate or ground-state pyrene dimerization under these conditions.

### Sample Deoxygenation

Initially, pyrene solutions were purged with N<sub>2</sub> for ca. 45 min prior to beginning the experiment. Unfortunately, possible pyrene decomposition was evidenced by a broad, red-shifted emission occurring at high temperatures. Further, on cooling these samples back to ambient conditions, we were unable to recover an emission spectrum that resembled pyrene in water prior to being subjected to supercritical conditions. In all subsequent experiments, the pyrene/water solutions were purged with Ar gas for approximately 45 min and were *then* subjected to multiple freeze-pump-thaw (FPT) cycles to remove all oxygen from the sample. The Ar/FPT technique resulted in no detectable decomposition of the pyrene at supercritical temperatures and recovery of a well-resolved pyrene emission

spectrum upon return of the system to ambient conditions.

### Instrumentation

A simplified schematic of our titanium high-pressure optical cell is shown in Figure 1. This cell is a modified version of a system used by the Brill group at the University of Delaware.<sup>52,53</sup> The cell is comprised of a titanium body (TB) which is coned and threaded at top and bottom to accept HiP (Erie, PA) high-pressure fittings (HPF). Sapphire windows (SW) (Insaco, Quakertown, PA) are sealed into each face of the cell using 24 K gold washers (GW) and titanium flanges (TF). This particular design allows the sapphire windows to be easily removed for cleaning. The same basic scheme has also been used to make cells with windows in 90° (not shown) and 180° (Figure 1) geometries for fluorescence and absorbance measurements, respectively. These cells have been tested to and routinely used for many hrs at 10,000 psia and a maximum temperature of 399 °C. These particular limits are set by the current pump and oven systems. Separate polyimide-coated fiber optics (FO) (CeramOptec, East Longmeadow, MA) are held against the sapphire optical windows using a special mounting flange that was developed in-house. This flange holds the optical fiber securely against the sapphire window face without breakage over the entire temperature range studied. The manufacturer specifications claim that these polyimide-coated fiber optics can routinely be operated up to 400°C.

For steady-state fluorescence measurements, the titanium high-pressure cell is incorporated into the experimental apparatus shown in Figure 2. A high-pressure syringe pump (P) (Isco, Model SFC-500) capable of producing up to 10,000 psia is operated in the constant-pressure mode to supply continuously oxygen-free pyrene/water solutions through 1/16" stainless steel tubing (with a preheater coil (PC)) to the high-pressure titanium cell (TC). The entire cell is located within a temperature-controlled GC oven (Hewlett Packard, Model 5730A). A valve and flow restrictor (FR) assembly are located outside the GC oven and each is adjusted during an experiment to maintain a constant solution flow (~ 300  $\mu\text{L}/\text{min}$ ) through the cell. This flow through approach is used to: (1) ensure that the fluid viewed within the cell is homogenous and supercritical and (2) minimizes the actual residence time of the solute at supercritical conditions. The system pressure is constantly monitored (0.03% accuracy) using a pressure transducer (PT) (Omega, Stamford, CT). The temperature is adjusted by the oven regulator and monitored using an insulated thermocouple (TH) (Simpson Accessories, Elgin, IL) located close to the Ti high-pressure cell within the oven. A He-Cd laser (HCL) (Omnichrome, Model 3074-20M) (325 nm) is used for excitation and an interference filter (10 nm FWHM, Oriel) is used to remove any extraneous plasma discharge from reaching the detection electronics. The laser beam is focused onto the proximal end of an optical fiber (FO), using a fused-silica lens (L) and XYZ translator (T), and the resulting fluorescence from the sample is collected using a second optical fiber whose output is collected and focused by a lens onto the entrance slit of an emission monochromator (M) (band pass = 2 nm). After proper wavelength selection, the signal is detected by a photomultiplier tube (D) and sent to a personal computer (PC) for processing. The remainder of the spectrofluorometer (SLM-AMINCO 48000 MHF) is configured in the standard ratiometric mode. Measurement of the pyrene  $I_1$  and  $I_3$  band intensities is made using software provided with the fluorometer.

The basic setup for our time-resolved fluorescence measurements is shown in Figure 3. A  $\text{N}_2$  laser (NL) (337 nm, LSI Incorporated, Model LS 337) operating at 20 Hz (with a pulse width of approximately 3 ns) is used for

excitation. A fused silica beam splitter (BS) serves to send a small portion of the excitation to a photodiode (PD) and the current pulse from the photodiode triggers a digital sampling oscilloscope (OSC) (Tektronics, Model TDS 350). The remainder of the excitation beam is focused onto the proximal end of an optical fiber, using a lens (L) and XYZ translator (T), and the resulting fluorescence emission from the sample is collected by a second optical fiber. A lens is used to collect and focus the emission through a bandpass filter (BPF) for wavelength selection onto the photocathode of a photomultiplier tube (PMT). The PMT dynode circuitry is designed for fast response and has been described in detail previously.<sup>54</sup> The pulsed output from the PMT is directed to the digital oscilloscope and commercial software provided with the oscilloscope is used to carry out all data acquisition. The excited-state fluorescence lifetime was determined from a linearized, logarithmic plot of the excited-state decay trace as described elsewhere.<sup>55,56</sup>

A large range of temperatures (from ambient to supercritical) and reduced densities were studied in this work. Special emphasis was placed on the reduced density ( $\rho_r$ ) range between 0.5-1.0 (where  $\rho_r = \rho_{\text{exp}}/\rho_c$ ) where a maxima in solvent-solute interaction is known to occur in other supercritical fluid systems.<sup>46</sup> Experiments were not conducted below  $\rho_r = 0.5$  because of poor signal-to-noise (S/N). The density and dielectric constant of water as a function of temperature and pressure were calculated using a simple, complete equation of state as described by Pitzer and Sterner.<sup>57</sup> All refractive index terms were calculated using the Clausius-Massotti equation and the appropriate molar refractivity and density.<sup>58</sup>

## Results and Discussion

### Steady-State Fluorescence Emission Studies

Figure 4 presents typical steady-state fluorescence spectra for pyrene in water at several temperatures using the apparatus described in Figure 2. For the low temperature spectrum, both  $I_1$  and  $I_3$  are well-resolved and easily distinguishable, showing that our fiber-optic setup provides adequate S/N and spectral resolution to measure the pyrene spectral features. Once we determined that our experimental apparatus was able to easily measure the pyrene emission spectrum, we began a systematic study, over a large range of temperatures and pressures, to determine the degree of thermal broadening and the extent of such on  $I_1/I_3$ . As the temperature is raised (Figure 4), spectral broadening of the pyrene emission occurs to the point that  $I_1$  and  $I_3$  become difficult to distinguish visually.

$I_1$  and  $I_3$  values are commonly measured by determining the intensity of the respective vibronic bands.<sup>43</sup> In liquid water at ambient conditions (25 °C), the  $I_1$  band occurs at 376 nm while the  $I_3$  band occurs at 383 nm (Figure 4). However, as temperature increases, the  $I_1$  and  $I_3$  bands clearly red shift *and* the entire spectral envelope broadens (Figure 4). Previous studies of pyrene in near-critical water<sup>42</sup> measured  $I_1$  and  $I_3$  at 376 and 383 nm, respectively; however, our results clearly illustrate that  $I_1$  and  $I_3$  bands actually red shift as the temperature increases. Therefore, it seems less than ideal to measure  $I_1$  and  $I_3$  at a constant wavelength under supercritical conditions in SCW. In order to circumvent these problems, we have carefully followed the spectral position of  $I_1$  and  $I_3$  (using the first derivative of the spectrum if needed) and calculate  $I_1/I_3$  using the actual band maxima.

Figure 5 presents the recovered  $I_1/I_3$  ratios, determined as previously described, for pyrene in water as a function of reduced density from 26.6 to 398.8 °C. Over this temperature range one can see that the  $I_1/I_3$  ratio is strongly

influenced by temperature, changing approximately 300% before “leveling off” below a reduced density of about two. It is well-known that temperature strongly influences  $I_1/I_3$  in polar liquids and supercritical  $\text{CO}_2$ ,<sup>44,45,50</sup> but measurements have not, to our knowledge, been made to the high temperatures explored in the current work. Therefore, our experimental data confirms the decrease in  $I_1/I_3$  with increasing temperature as previously observed, but over a much higher temperature range. Our results also suggest that this temperature-induced decrease in  $I_1/I_3$  for pyrene in water does not affect the data collected *above* about 280 °C.

To provide quantitative information on the local environment surrounding pyrene in SCW, we require a link between the experimental measurables ( $I_1/I_3$ ) and some physicochemical property of the solvent. Fortunately, pyrene has been used previously to determine the extent of local density augmentation in supercritical  $\text{CO}_2$ .<sup>46,47</sup> For pyrene in normal liquid solvents,  $I_1/I_3$  is linear with the well-known  $\pi^*$  polarity scale, where  $a$  and  $b$  are constants.<sup>46,47,59</sup>

$$I_1/I_3 = a + b\pi^* \quad (1)$$

The  $\pi^*$  term, in turn, has been shown to depend linearly on the dielectric cross term,  $f(\epsilon, n^2)$  where  $c$  and  $d$  are constants,  $\epsilon$  is the solvent dielectric constant, and  $n$  is the solvent refractive index.<sup>46,47</sup>

$$\pi^* = c + df(\epsilon, n^2) \quad (2)$$

$$f(\epsilon, n^2) = [(\epsilon - 1)/(2\epsilon + 1)][(n^2 - 1)/(2n^2 + 1)] \quad (3)$$

Therefore, the relationship between  $I_1/I_3$  and the dielectric cross term should be linear, in the absence of any local density effects.<sup>46,47</sup>

$$I_1/I_3 = A + Bf(\epsilon, n^2) \quad (4)$$

where  $A$  is the vapor phase  $I_1/I_3$  value for pyrene (0.41)<sup>43</sup> and  $B$ , the slope, is determined by the pyrene  $I_1/I_3$  at high density, liquid-like fluid values.

Figure 6 presents our  $I_1/I_3$  data as a function of  $f(\epsilon, n^2)$  at 379.9 °C. The solid and dashed lines are the values predicted if there were no solute-fluid clustering. The line labeled “ $T_{\text{uncomp}}$ ” is derived by using the  $I_1/I_3$  at the highest liquid densities *without* any compensation for the well-known decrease in  $I_1/I_3$  with temperature.<sup>44,45,50</sup> The dashed “ $T_{\text{comp}}$ ” line attempts to take into better account the decrease in  $I_1/I_3$  with increasing temperature by using high temperature, high density  $I_1/I_3$  values. The upward deviation of the experimentally determined  $I_1/I_3$  with respect to the “ $T_{\text{comp}}$ ” line is indicative of solute-fluid clustering<sup>46,47</sup> between pyrene and SCW. Using the differences between the experimental  $I_1/I_3$  values and the predicted  $I_1/I_3$  values, we can estimate the local water density immediately surrounding

the average pyrene molecule as a function of bulk fluid density.<sup>46,47</sup> By dividing this local term by the bulk SCW density ( $\rho_{\text{local}}/\rho_{\text{bulk}}$ ) we can determine the extent of local density augmentation (clustering) surrounding the pyrene.

Figure 7 illustrates the  $\rho_{\text{local}}/\rho_{\text{bulk}}$  as a function of reduced density for pyrene in SCW at 379.9, 384.4, and 398.8 °C. These results show several interesting points. First, the largest degree of density augmentation occurs well below the critical density. Second, the degree of density augmentation is on the order of 500% at a reduced density near 0.5. Finally, we see that as the temperature is increased, there is an apparent decrease in the degree of fluid density augmentation occurring around the pyrene molecule. This decrease in solute-fluid interactions with increasing temperature has been previously shown to occur in other, more mild supercritical fluid systems as well.<sup>49,60</sup>

It is known, from work on other supercritical fluids, that a maximum in solute-fluid clustering occurs at approximately one-half the critical density.<sup>46,47</sup> Although it is difficult to make measurements at these low fluid densities in SCW, we also observe data fully consistent with a maximum in clustering occurring in this region for pyrene in SCW. Thus, there is no evidence for solute-fluid rarefaction in this system. In supercritical CO<sub>2</sub>, the typical, maximal local density enhancement is on the order of 230% for pyrene.<sup>46</sup> Clearly, we observe a much greater degree of local density enhancement in SCW. This is likely due in part to the ability of water to hydrogen bond more with itself and/or due to the more polarizable nature of water relative to CO<sub>2</sub>.

#### Time-Resolved Fluorescence Studies

Time-resolved fluorescence allows one to access processes that occur on a time scale similar to the excited-state fluorescence lifetime.<sup>55,56</sup> Figure 8 presents a typical series of time-resolved fluorescence decay traces for pyrene in SCW at several temperatures and pressures. Each of these fluorescence decays, within the current time resolution of the apparatus (5-6 ns), is well-described by a monoexponential decay. Fluorescence lifetimes were not measured in the low density region because of inadequate S/N. Figure 9 shows that as temperature is increased, we see a systematic decrease in the pyrene lifetime in SCW. This decrease correlates well with the  $I_1/I_3$  data (Figure 5). Analysis of these data in terms of an Arrhenius relationship, allows one to determine if the decrease in fluorescence lifetime is due to a change in solvation and relaxation pathways or if the decrease is simply a thermally-activated process. Examination of the Arrhenius plot (Figure 9, inset) shows that the systematic decrease in lifetime is exponentially activated.

Inspection of the literature shows that this same type of phenomena has been previously observed for pyrene in ethanol and liquid paraffin between -100 to 140 °C.<sup>61</sup> The activation energy for this process in ethanol and paraffin were determined to be 9.49 and 12.97 kJ/mol, respectively. For the current work, the recovered activation energy in water was determined to be  $4.07 \pm 0.1$  kJ/mol. Although differences in these activation energies can not be easily explained, it is interesting to note that the activation energy for the process decreases as the solvent polarity increases. To determine the origin of the decrease in fluorescence lifetime with increasing temperature, it is useful to compare the recovered activation energies with the activation energy predicted for simple viscous flow conditions ( $E_\eta(\text{ethanol}) \sim 3$  kJ/mol;<sup>61</sup>  $E_\eta(\text{paraffin}) \sim 50$  kJ/mol;<sup>61</sup>  $E_\eta(\text{H}_2\text{O}) = 15.5$  kJ/mol<sup>62</sup>). Because the recovered activation energy for the decrease in pyrene fluorescence lifetime is different than the activation energy for viscous flow, it can be concluded that this decrease is not due to quenching of the pyrene lifetime by diffusion of an unknown species.<sup>61</sup>

In order to explain this decrease in lifetime with increasing temperature, it is useful to examine the electronic states of pyrene more closely. Specifically, pyrene is known to have a triplet state ( $T_2$ ) that lies only 300  $\text{cm}^{-1}$  above the first excited singlet state ( $S_1$ ).<sup>63</sup> Under normal conditions, intersystem crossing to this triplet state does not occur due to a large activation energy between these states, causing the crossing to be energetically unfavorable.<sup>61,63</sup> Further, if the rate of vibrational relaxation is fast compared to the rate of intersystem crossing ( $k_{ISC}$ ) between  $S_1$  and  $T_2$ , the rate of intersystem crossing can be defined by:<sup>64</sup>

$$k_{ISC} = (4\pi^2 D/h) J^2 F(E) \quad (5)$$

where  $D$  is the density of the final states,  $J$  is the electronic transition matrix element, and  $F(E)$  is the Franck-Condon factor summed over all states.

As the system temperature is increased, there is a concomitant increase in the population of pyrene molecules within the upper vibrational levels (denoted  $v$ ) of the ground-state manifold ( $S_0$ ). On excitation, these molecules, within the upper vibrational levels of  $S_0$ , become promoted into the  $S_1$  manifold. By applying the Boltzmann distribution, one can write the intersystem crossing rate as:<sup>64</sup>

$$k_{ISC} = \frac{\sum_{E(v)=E}^{\infty} k_{ISC}(v) e^{-E(v)/RT}}{\sum_{E(v)=0}^{\infty} e^{-E(v)/RT}} \quad (6)$$

where  $k_{ISC}(v)$  and  $E(v)$  are the rate of intersystem crossing and excess vibrational energy, respectively, at a particular vibrational level within the  $S_1$  manifold. Therefore, as the upper vibrational level "gateways" within  $S_1$  become populated, the rate of intersystem crossing to the triplet state becomes thermally activated, causing an increase of intersystem crossing to the triplet state. This leads to an increased nonradiative pathway, and subsequent decrease in the excited-state pyrene fluorescence lifetime in water at higher temperatures. Thus, although solute-fluid clustering is substantial within the pyrene/SCW system, it does not appear to affect the pyrene emissive rates beyond those seen due to thermal activation.

### Conclusions

Although significant spectral broadening occurs in the pyrene emission spectra with increasing temperature, we were able to estimate  $I_1/I_3$  ratios as a function of reduced density over a wide range of temperatures.  $I_1/I_3$  is seen to systematically decrease (approximately 300%) with increasing temperature before leveling off above  $\sim 280^\circ\text{C}$ . This phenomenon has little to do with solvation per se, but is a result of temperature-induced changes in the pyrene emission.<sup>44,45,50</sup> Deviation of the experimentally determined  $I_1/I_3$  values from the predicted values above the critical temperature allows us to estimate the extent of local density augmentation surrounding pyrene dissolved in SCW. These



results show that near one half the fluid critical density, the local density augmentation is on the order of five times greater than the bulk water density. This degree of solute-fluid clustering decreases as the system temperature and pressure are increased. In fact, at a reduced density of 1.2 - 1.3, the local and bulk SCW densities appear comparable surrounding pyrene.

Time-resolved fluorescence measurements of pyrene in water show that as the temperature of the system is increased, a subsequent decrease in the pyrene fluorescence lifetime occurs. This decrease is shown to be exponentially activated and is a manifestation of thermally populating upper vibrational levels within  $S_0$  and in turn  $S_1$  that open up promoter modes to a nearby triplet state ( $T_2$ ). Proximity to the solvent critical point does not apparently affect the pyrene fluorescence decay kinetics. Thus, while solute-fluid clustering is clearly evident, it does not apparently influence the pyrene photophysics.

### **Probing Intermolecular Interactions in Supercritical Alkanes: Mock Supercritical Aviation Fuels**

With the eventual introduction of advanced high-speed aircraft, the onboard fuel supply will likely be called upon to cool the plane fuselage. In some instances, the circulating fuel may be heated such that it becomes a supercritical fluid (*vide supra*). It is therefore important to question how the environment surrounding a dissolved solute/additive may change in an aviation fuel raised above its critical point as a function of fuel density, temperature, and pressure. Clearly, such will govern not only fuel performance and lifetime, but also the design and overall lifetime of aircraft components.

As a step toward addressing this issue, we have used static fluorescence spectroscopy to determine the effects of supercritical temperatures on several simple alkane fuel precursors (i.e., *n*-pentane, *n*-hexane and *n*-heptane). Fluorescence spectroscopy provides an ideal tool to quantify molecular-level interactions occurring in supercritical fluids.<sup>65</sup> Due to the inherently low detection limits of fluorescence measurements,<sup>56</sup> it is possible to work at "infinite dilution," thus minimizing any solute-solute interactions. By using a solute/probe that is sensitive to its local environment, we are able to access directly the local environment surrounding the probe.

We have used the fluorescent additive/probe pyrene to quantify local intermolecular interactions and to determine how these interactions may differ from the bulk fluid properties. Pyrene is an ideal solute because its photophysics are well-known and it has been thoroughly studied in a variety of media,<sup>43-50</sup> including other supercritical fluids.<sup>46-50</sup> However, there are many special experimental considerations that must be taken into account when making these measurements in supercritical alkanes. Due to the high temperature needed to generate a supercritical alkane ( $T_c$  (pentane) = 196.6 °C;  $T_c$  (hexane) = 234.4 °C;  $T_c$  (heptane) = 267.30 °C), the solute must be stable at harsh temperatures with minimal spectral broadening. Work in this laboratory (*vide supra*) has shown that pyrene is stable up to 400 °C in supercritical water with measurable  $I_1/I_3$  ratios,<sup>66</sup> and is therefore an ideal probe for the supercritical alkane work. Deoxygenation is also imperative for these experiments, not only for the stability of the solute, but also to avoid combustion of the alkanes at high temperatures and pressures.

## Experimental

Pyrene (99.9%) was purchased from Aldrich and used as received. Spectrophotometric grade (99%) *n*-pentane, *n*-hexane and *n*-heptane were purchased from Aldrich and used without further purification. Solutions of pyrene in each of the respective alkanes were prepared by first pipetting the appropriate amount of pyrene/ethanol stock solution into a vessel. The ethanol was then evaporated off, and alkane was added to the vessel to make a 10  $\mu$ M solution. All pyrene/*n*-alkane solutions are deoxygenated by purging with Ar for approximately 1 hr prior to the experiment.

Steady-state fluorescence measurements are made using the titanium high-pressure cell and fiber optic setup described previously (*vide supra*).<sup>66</sup> Deoxygenated pyrene/alkane solution is continuously flowed through the cell using a flow restrictor assembly and high-pressure syringe pump operated in constant pressure mode (*vide supra*). A He-Cd laser (Omnichrome, Model 3074-20M) (325 nm) is used for excitation and an interference filter (10 nm FWHM, Oriel) is used to remove any extraneous plasma discharge from reaching the detection electronics. All measurements are made using the SLM-AMINCO 48000 MHF spectrofluorometer configured in the standard ratiometric mode. Measurement of the pyrene  $I_1$  and  $I_3$  band intensities is made using software provided with the fluorometer.

A large range of reduced densities at a reduced temperature ( $T_r$ ) of 1.01 (where  $T_r = T_{\text{exp}}/T_c$ ;  $T_{\text{exp}}$  = experimental temperature;  $T_c$  = critical temperature) were studied for each of the *n*-alkanes with emphasis placed on the reduced density range between 0.5-1.0 where maximum solvent-solute interactions are known to occur in other supercritical fluid systems.<sup>46</sup> The density of each of the alkanes was estimated as a function of temperature and pressure using a commercial software package (SFSolver, Isco, Inc). All refractive index and dielectric constant terms were calculated using the Clausius-Massotti equation and the appropriate molar refractivity and density.<sup>58</sup>

## Results and Discussion

The pyrene  $I_1/I_3$  ratio has been used to provide insight into the nature of solute-fluid interactions in supercritical water<sup>66</sup> and CO<sub>2</sub>.<sup>46,47</sup> A similar format has been used to quantify the intermolecular interactions occurring in supercritical alkane systems. Figure 10 presents  $I_1/I_3$  ratios as a function of reduced density for pyrene in *n*-pentane (Panel A), *n*-hexane (Panel B), and *n*-heptane (Panel C) at  $T_r = 1.01$ . From these data, equations 1 through 4 were used to relate the  $I_1/I_3$  ratios to the physical properties of the fluid via the dielectric cross term ( $f(\epsilon, n^2)$ ). Figure 11 shows  $I_1/I_3$  ratios as a function of the dielectric cross term for pyrene in *n*-pentane (Panel A), *n*-hexane (Panel B), and *n*-heptane (Panel C) with associated uncertainties. The solid lines each represent the theoretical  $I_1/I_3$  ratios in the absence of any solute-solvent interactions. Upward deviation from this line is indicative of solute-fluid clustering, or an increase in the local density surrounding the pyrene molecule.<sup>46,47,66</sup> The upward deviation of the experimental  $I_1/I_3$  ratios relative to the theoretical line (no clustering) is then used to calculate the local alkane density surrounding pyrene in each of these alkane systems. Figure 12 presents the calculated local alkane density ( $\rho_{\text{local}}$ ) divided by the bulk alkane density ( $\rho_{\text{bulk}}$ ) in each of the *n*-alkanes. The  $\rho_{\text{local}}/\rho_{\text{bulk}}$  term can be used to estimate the degree of solute-fluid interactions occurring surrounding pyrene<sup>46,47,66</sup> in these supercritical alkanes. Several things are evident from inspection of Figure 12. First, the maximum degree of fluid clustering occurs at around one-half the critical density. Second, the maximum in local

density is 3 to 4 times the bulk alkane density. Finally, the degree of clustering is similar in each of the alkane systems studied.

It has been previously shown for other supercritical fluids (e.g., CO<sub>2</sub>, H<sub>2</sub>O) that a maximum in solute-fluid clustering occurs at approximately one-half the critical density.<sup>46,47,66</sup> This is fully consistent with our observations in all supercritical alkanes, although measurements have not yet been made in the very low density region. The maximum in local density in the alkane systems of approximately 3 to 4 times the bulk density is consistent with the results seen in other fluids<sup>46,66</sup> (*vide supra*) although slightly higher than those observed in supercritical CO<sub>2</sub>.

### Conclusions

An increase in local alkane density surrounding pyrene dissolved in supercritical C<sub>5</sub>, C<sub>6</sub>, and C<sub>7</sub> *n*-alkanes is observed at one half the critical density. The degree of augmentation (3 to 4 times  $\rho_{\text{bulk}}$ ) and the maximum in clustering, at one-half the critical density, is fully consistent with previous results in other supercritical fluids.<sup>46,47,66</sup> To our knowledge, this type of clustering phenomena has not been previously observed within supercritical alkane systems. This work points out that additional experimentation is needed to understand how these solute-solvent interactions in aviation fuel precursors will affect fuel performance.

### Effects of CO<sub>2</sub> Sorption on Molten Polymers Dynamics

Poly(dimethylsiloxane) (PDMS) polymers are unique silicone polymers which possess a very low glass phase transition temperature ( $T_g$ ) of  $\sim 150\text{K}$ .<sup>67,68</sup> For this reason, PDMS polymers are molten viscous polymers which exhibit flow characteristics at ambient temperatures and are therefore model polymers for other systems with much higher processing temperatures.<sup>67,68</sup> PDMS polymers have found a wide range of applications from coatings and implants to hydraulic fluids.<sup>67,68</sup>

There is significant interest in developing simple methodologies to control the transport properties of solid or molten polymers because such governs aspects of polymer synthesis, polymer reactivity, and polymer processing.<sup>2,67,68</sup> It is well-known that CO<sub>2</sub> can be used to alter the characteristics of certain polymer systems.<sup>69-72</sup> For example, recent work has shown that molten polymers like PDMS can sorb tremendous amounts of CO<sub>2</sub> leading to a dramatic decrease in the polymer bulk viscosity.<sup>69-72</sup> However, there is much to be learned about how gas effusion affects polymer dynamics, polymer free volume, the mobility of dissolved solutes, and polymer transport properties.

Rotational reorientation measurements have been used extensively to determine the nature of interactions occurring within both amorphous polymer matrices,<sup>73-76</sup> and at elastic cross-link junctions within various polymer networks.<sup>77-80</sup> For example, Fayer and co-workers<sup>76</sup> have recently used the rotational reorientation dynamics of dansylamide attached to a trifunctional silane within PDMS melts to attempt to relate local rotational dynamics to bulk polymer properties. However, the rotation of the rather small ( $\sim 3\text{\AA}$  in length) probe did not correlate well with bulk polymer dynamics.

We have used a large, neutral solute in conjunction with time-resolved fluorescence anisotropy techniques to

correlate the rotational motion of the solute to bulk polymer properties. Specifically, we are interested in the effects of the addition of CO<sub>2</sub> (at both ambient and supercritical conditions) on the behavior of model dopants within the polymer matrix and whether we can track such effects. The model solute that we have chosen for our study is BTBP, which has been previously used for rotational reorientation measurements within a variety of media.<sup>81-83</sup> BTBP is a large (28 Å in length) neutral solute. Thus, there will be no charge interactions with the polymer matrix. BTBP has a unity quantum yield,<sup>81-83</sup> therefore small amounts may be dispersed within the polymer while maintaining sufficient S/N. BTBP has been reported to be a spherical rotor with a single rotational correlation time.<sup>81-83</sup> For this work, we have measured the effects of polymer molecular weight and CO<sub>2</sub> on the rotational dynamics of BTBP to correlate the rotational correlation time with the polymer's properties.

## Experimental

### Preparation of Bulk Polymer Solutions

A broad average molecular weight range (MW = 1250, 2000, 3780, 5970, 9430, 13650, 28000, and 49350 g/mol) of methyl-terminated PDMS was purchased from United Chemical Technologies, Inc (Bristol, PA) and used without further purification. *N,N'*-Bis(2,5-di-*tert*-butylphenyl)-3,4,9,10-perylenedicarboximide (BTBP) was purchased from Aldrich and used as received.

Stock solutions of BTBP (1 mM) were prepared in absolute ethanol. BTBP is randomly dispersed within the PDMS via the following protocol: 1) the appropriate amount of BTBP stock solution is micropipetted into a vial; 2) the vial is then placed within a hot oven for approximately 1 hr to evaporate any remaining ethanol solvent; 3) after cooling, the appropriate quantity of PDMS is added to make a final solution that is 1 μM BTBP; and 4) the solutions are stirred (with gentle heating for higher molecular weight samples) for approximately 2 wks to thoroughly disperse the BTBP throughout the polymer. There is no evidence of BTBP aggregates.

### Addition of CO<sub>2</sub> to the Polymer Samples

CO<sub>2</sub> is added to the polymer solutions via a syringe pump assembly which continuously delivers CO<sub>2</sub> to a high-pressure cell containing the polymer sample. The stainless steel high-pressure cell was developed in-house and has been described in detail previously.<sup>84</sup> The cell has an optical pathlength of approximately 1 cm and contains quartz optical windows (Behm Quartz Industries, Dayton, OH) which have been previously shown to exhibit no detectable pressure induced birefringence over the pressure range studied.<sup>81</sup>

The BTBP/PDMS solution (3.75 mL) is directly pipetted into the high-pressure cell (internal volume = 5mL) into which a teflon-coated stir bar has been placed. A valve assembly is then connected to the high-pressure cell which attaches the cell to a high pressure syringe pump (Isco, Model 260D, Lincoln, NE) operating in constant pressure mode. Throughout the experiment, a Haake A80 temperature bath is used for temperature control. The temperature is monitored using a solid-state thermocouple (Cole Parmer, Vernon Hills, IL) and pressure is monitored within ±1 psi using a calibrated Heise pressure gauge.

The cell is first charged to the highest pressure (~2500 psi) and allowed to equilibrate at experimental temperature with constant stirring. Initially, polymer samples were equilibrated overnight before rotational reorientation

measurements were made. Figure 13 presents the recovered rotational correlation time for BTBP in PDMS (MW = 28000 g/mol) as a function of time at approximately 1000 psia. Time zero corresponds to 12 hrs of equilibration time with stirring. We observe an initial increase in the rotational correlation time which we attribute to increased pressure within the system from delivery of the CO<sub>2</sub>. The systematic decrease of the rotational correlation time with time tracks the diffusion of CO<sub>2</sub> into the PDMS and subsequent dilation of the polymer. This systematic decrease due to continued dilation of the polymer was shown to continue for upwards of 1 wk. Therefore, all subsequent PDMS/CO<sub>2</sub> samples were equilibrated for approximately 2 wks with stirring at the initial experimental temperature and pressure.

### Instrumentation

All time-resolved measurements were made using a multiharmonic frequency-domain fluorometer (SLM-AMINCO 48000 MHF). For steady-state measurements, a Xe arc lamp is used for excitation with a monochromator for appropriate wavelength selection of emission and excitation (bandpass = 4 nm). The 514.5 nm line of a CW Ar<sup>+</sup> laser (Coherent, Model Innova 400) is used for excitation during all time-resolved experiments. The output from the laser is passed through an interference filter to eliminate any extraneous plasma discharge from reaching the detector.

Fluorescence from the sample is monitored through a 550 longpass filter and a polarizer set at the magic angle condition for fluorescence lifetime measurements.<sup>85</sup> Sinusoidally modulated light is generated using a Pockels cell driven at 5 MHz, and data is collected from 5 to 200 MHz (39 frequencies). At least 9 replicate measurements were made. Rhodamine 6G in water was used as the reference lifetime for all excited-state fluorescence lifetime measurements ( $\tau = 3.85$  ns).<sup>86</sup> The BTBP fluorescence lifetime was found to be constant over the range of experimental conditions. Operation of a typical MHF frequency-domain fluorometer has been described in great detail elsewhere.<sup>87-89</sup> Phase and demodulation data were fit to various test models by using a commercially available global analysis software package (Globals Unlimited).<sup>90</sup>

Frequency-domain measurements of the time-resolved decay of anisotropy are made by measurement of the differential phase angle ( $\Delta = \theta_{\perp} - \theta_{\parallel}$ ) and polarized modulation amplitude ( $\Lambda = AC_{\parallel}/AC_{\perp}$ ). The decay of the intensity of the parallel ( $I_{\parallel}(t)$ ) and perpendicular ( $I_{\perp}(t)$ ) components of the polarized fluorescence in the frequency domain may be described by:<sup>89,91</sup>

$$I_{\parallel}(t) = 1/3[I(t)(1+2r(t))] \quad (7)$$

$$I_{\perp}(t) = 1/3[I(t)(1-r(t))] \quad (8)$$

where  $r(t)$  is the fluorescence decay of anisotropy. Assuming that the fluorophore is a spherical rotor, the decay of anisotropy can be written:

$$r(t) = r_0 \exp\left(-\frac{t}{\phi}\right) \quad (9)$$

where  $r_0$  is the limiting anisotropy, the anisotropy measured in the absence of rotational motion. The rotational correlation time,  $\phi$ , may then be recovered by fitting the differential phase angle ( $\Delta$ ) and the polarized modulation ratio ( $\Lambda$ ) as a function of frequency using a non-linear least squares program:<sup>91</sup>

$$\Delta = \arctan\left[\frac{D_{\parallel}N_{\perp} - D_{\perp}N_{\parallel}}{N_{\parallel}N_{\perp} + D_{\parallel}D_{\perp}}\right] \quad (10)$$

$$\Lambda = \left[\frac{N_{\parallel}^2 + D_{\parallel}^2}{N_{\perp}^2 + D_{\perp}^2}\right]^{1/2} \quad (11)$$

where N and D are the polarized components of the sine and cosine Fourier transform, respectively. The rotational correlation time is then fit using a non-linear least squares minimization of the chi-squared ( $\chi^2$ ) parameter as defined by:<sup>91</sup>

$$\chi^2 = \frac{1}{D} \sum_{\omega} \left( \frac{\Delta_m(\omega) - \Delta_c(\omega)}{\sigma_{\Delta}} \right)^2 + \frac{1}{D} \sum_{\omega} \left( \frac{\Lambda_m(\omega) - \Lambda_c(\omega)}{\sigma_{\Lambda}} \right)^2 \quad (12)$$

where the subscripts c and m denote the computed and measured differential phase angles and polarized modulation ratios, respectively,  $\sigma$  is the variance, and D is the number of degrees of freedom. The goodness of fit between the model and the experimental data is determined by the closeness of the  $\chi^2$  parameter to unity as well as the randomness of the residuals around zero.

## Results and Discussion

### Rotational Reorientation Dynamics within Molten Polymers

The determination of the rotational correlation time of BTBP as a function of MW allows us to determine if our large neutral solute (BTBP) can track the entanglement value ( $M_e$ ) for PDMS. The entanglement value is a characteristic value of amorphous polymers in which the chains within the polymer become too long to slip past one another easily.<sup>67,68</sup> Although the polymer will still exhibit flow characteristics as the MW increases, a “leveling off” of physicochemical properties (i.e., viscosity, refractive index) occurs past the entanglement value.<sup>67,68</sup> Subsequently, if BTBP is large enough to fill the free volume fully within the polymer matrix, the rotational correlation time should level off above the entanglement value.

Figure 14 presents typical differential phase angle (Panel A) and polarized modulation ratio (Panel B) data for

BTBP in several PDMS molecular weight polymers. The points represent actual experimental data and the lines are recovered best fits to a single exponential decay law. From this data, it is obvious that as the molecular weight of the polymer increases, the differential phase angle and polarized modulation ratio subsequently increase, indicating an increase in the rotational correlation time. Figure 15 presents the recovered rotational correlation times as a function of PDMS molecular weight. We see that as the polymer molecular weight increases, there is an initial increase in the rotational correlation time and a leveling off at  $\sim 10000$  g/mol. Entanglement values have been reported for PDMS polymers in the literature ranging between  $\sim 8000^{92}$  and  $8625^{67}$  g/mol. Our rotational reorientation data “breaks” at a value that is near these reported literature values. Thus, the large size of the BTBP probe allows us to easily track the PDMS chain entanglement process.

Rotational reorientation data as a function of PDMS molecular weight also allows us to establish a convenient link between  $\phi$  and some aspect of the polymer. Figure 16 presents the BTBP rotational correlation time as a function of polymer density. The solid line is the straight line fit. This will be used later to correlate the recovered rotational correlation times in PDMS diluted with  $\text{CO}_2$ .

#### Effect of $\text{CO}_2$ on PDMS Dynamics

Figure 17 presents the recovered rotational correlation time of BTBP as a function of added  $\text{CO}_2$  pressure in several representative molecular weight PDMS polymers (i.e., 1250, 9430, 13650, and 28000 g/mol) at 25 °C. There are several aspects of this data that merit special attention. First, the rotational correlation time decreases dramatically (up to 5 times) with the addition of liquid  $\text{CO}_2$  from the rotational correlation times observed in the neat PDMS melts. It is known that  $\text{CO}_2$  can swell PDMS up to 50% by weight,<sup>69-72</sup> and our data is fully consistent with the dilation and subsequent decrease in bulk density of the PDMS polymer with addition of  $\text{CO}_2$ . Second, it appears that the decrease in rotational correlation time can be further decreased by increasing the  $\text{CO}_2$  pressure. We observe that the rotational correlation time “snaps” to a much lower value between 500 and 1000 psi of  $\text{CO}_2$  for all molecular weight polymers before leveling off with increasing pressure above 1000 psi. Examination of the phase equilibria for  $\text{CO}_2$  at 25 °C shows that the transition between the gas and liquid state of  $\text{CO}_2$  occurs between these pressures. The higher density liquid  $\text{CO}_2$  dilates and swells the polymer to a greater extent than the less dense gaseous  $\text{CO}_2$ . The density of the swollen polymer may be estimated by using the relationship between the PDMS density and the recovered rotational correlation time in neat PDMS melts as a function of molecular weight (Figure 16). Figure 18 presents the calculated polymer densities as a function of added  $\text{CO}_2$  for each of the polymers.

These results prompted us to question whether the BTBP rotational correlation time could actually be tuned by addition of  $\text{CO}_2$  to the polymer. From the previous 25 °C experiments, we noted a sharp change in the rotational correlation time as a function of pressure. It is known that  $\text{CO}_2$  above its critical temperature ( $T_c = 31.1$  °C;  $P_c = 1070.4$  psia) exhibits no phase boundary with increasing pressure. Thus, we questioned if we could tune the BTBP rotational reorientation time. Figure 19 presents the recovered rotational correlation times for BTBP in PDMS (MW = 9430 g/mol) at ambient ( $T = 25.0$  °C) and supercritical ( $T = 36.5$  °C) conditions as a function of  $\text{CO}_2$  pressure. The recovered rotational times for BTBP in the PDMS swelled with supercritical  $\text{CO}_2$  are slightly lower due to the increase in

temperature which subsequently decreases  $\phi$  (Figure 19A). More interestingly, closer examination of the  $T = 36.5\text{ }^{\circ}\text{C}$  data in Figure 19B shows that instead of the sharp change in rotational correlation time observed with increasing  $\text{CO}_2$  pressure in the ambient system (*vide supra*), we instead observe a very gradual easily tuned decrease in the BTBP rotational correlation time. The inherent tunability of the physicochemical properties of supercritical  $\text{CO}_2$  with temperature and pressure can therefore be used to tune the PDMS matrix and hence the BTBP rotational dynamics.

### Conclusions

We have shown that our large model solute, BTBP, can be used to track the bulk polymer dynamics within the PDMS polymer matrix. The rotational correlation time is shown to scale with polymer molecular weight before leveling off above the known PDMS entanglement value. The rotational correlation time of BTBP within these molten polymers has been used to relate the recovered rotational dynamics within  $\text{CO}_2$ -diluted PDMS to the bulk density of the swelled polymer matrix. We have shown that the addition of ambient temperature  $\text{CO}_2$  to PDMS causes a dramatic decrease in the BTBP rotational correlation time near the  $\text{CO}_2$  gas-liquid phase boundary ( $25\text{ }^{\circ}\text{C}$ ). Addition of supercritical  $\text{CO}_2$  allows us to tune the bulk polymer density and the BTBP rotational dynamics.

### Cosolvent Effects on Rotational Reorientation Dynamics in Supercritical $\text{CO}_2$

The inherent tunability of supercritical fluids have made them attractive for use in separations,<sup>1-4</sup> chemical reactions,<sup>6-13</sup> and extractions.<sup>5</sup> Supercritical  $\text{CO}_2$  ( $\text{scCO}_2$ ) is environmentally friendly, inexpensive, and has very mild critical parameters ( $T_c = 31.1\text{ }^{\circ}\text{C}$ ;  $P_c = 1070.4\text{ psia}$ ;  $\rho_c = 0.468\text{ g/mL}$ ). Supercritical  $\text{CO}_2$  has found a wide range of industrial applications including replacement of hazardous solvents in the clothing dry cleaning process, decaffeination of coffee beans, and a wide range of extractions and reactions in the petroleum, polymer and pharmaceutical industry.<sup>2-4</sup> Although  $\text{scCO}_2$  is by far the most commonly used supercritical solvent, it is, unfortunately, a rather poor solvent for polar solutes.  $\text{CO}_2$  is nonpolar and has a relatively low solvent strength. In addition, although there is inherent tunability of the  $\text{scCO}_2$  solvent strength with pressure and temperature, it still exhibits a cohesive energy density less than cyclohexane and orders of magnitude less than common industrial solvents such as methanol and acetonitrile. This low cohesive energy density translates into a lack of selectivity for chromatographic separations, extractions and reactions.

There are several approaches to “modify”  $\text{CO}_2$  in order to improve the power and hence selectivity of the solvent. The addition of small quantities ( $x = 0.1 - 5\text{ mole\%}$ ) of an organic entrainer (cosolvent) have been shown to dramatically increase solute loading as well as increase reaction/separation selectivity.<sup>70,93-95</sup> More recently, a perfluoropolyether-based surfactant (PFPE) has been used to form stable reverse micelles in  $\text{scCO}_2$ .<sup>96</sup> However, although promising, these PFPE microemulsions are a relatively new technology and have not yet been thoroughly characterized. In addition, there are many recognized advantages to using the more simple cosolvent-modified systems. For example, cosolvents do not significantly alter the  $\text{CO}_2$  critical properties. In addition, only small quantities of entrainer need to be added, giving both an economical and a waste disposal advantage. Enhanced solute loading of both polar and nonpolar solutes is achieved as well as improving selectivity of reactions, extractions, and separations.<sup>70,93-95</sup>



Many of these desirable properties have been attributed to preferential solvation of the cosolvent around the dissolved solute in these supercritical systems. Given this, the goal of this work was to experimentally quantify preferential solvation in a supercritical fluid system.

Rotational reorientation dynamics measurements offer a convenient means to relate the solute's dynamics to the local solvent microdomain (*vide supra*). We have used the fluorescent probe BTBP in concert with time-resolved fluorescence anisotropy techniques to determine rotational dynamics of a solute as a function of MeOH-modified scCO<sub>2</sub> density. These data have allowed us to quantify the extent of solute-fluid clustering and determine how the solvation shell surrounding the solute changes as a function of bulk density.

## Experimental

### Preparation of the Cosolvent Mixture

Methanol (spectrophotometric grade, Aldrich) is charged into a stainless steel vessel equipped with two 1/8" tubing pieces, one leading to a SFC grade CO<sub>2</sub> tank (Scott Specialty Gases, Plumsteadville, PA) and the other to a high-pressure syringe pump (Model 260D, Isco). After charging the MeOH, the vessel is promptly placed into an ice bath and CO<sub>2</sub> is added to the vessel as it is vigorously shaken and stirred to ensure that the mixture is one phase. The mixture is then transferred into the syringe pump, and the pump piston is run in and out several times to ensure mixing. The pump head is then heated to experimental temperature (~45 °C) and allowed to equilibrate for several days to complete the mixing process. The MeOH/CO<sub>2</sub> mixture is then delivered to our high-pressure stainless steel optical cell (*vide supra*) which has been previously charged with BTBP. All MeOH/CO<sub>2</sub> mixtures for this experiment were 5 mol% MeOH. The cell is initially charged with the highest experimental pressure of the cosolvent mixture (~3000 psia) and the pressure is gradually decreased throughout the experiment until the mixture becomes biphasic.

All mixture densities and viscosities were calculated using the equations described by Foster and co-workers.<sup>97</sup>

### Results and Discussion

The steady-state emission and excitation spectra of BTBP in MeOH-modified CO<sub>2</sub> at 45°C and 2667 psia are shown in Figure 20. Figure 21 presents the effects of density on the excited-state fluorescence lifetimes for BTBP in MeOH-modified CO<sub>2</sub> at 46.0 °C. As mentioned previously, it has been established that the fluorescence lifetime for BTBP is relatively constant as a function of its local environment.<sup>82,83</sup> However, we observe a systematic decrease in the BTBP fluorescence lifetime with increasing MeOH/CO<sub>2</sub> bulk density. This phenomena may be readily explained by the Strickler-Berg relationship, which links the radiative rate of the fluorophore ( $k_r$ ) to the properties of the solvent and fluorophore through the following equation:<sup>63</sup>

$$k_r = 2900n^2\nu_0^2 \int \epsilon d\nu \quad (13)$$

where  $n$  is the solvent refractive index,  $\nu_0$  is the peak frequency in the fluorophore absorbance spectrum, and  $\int \epsilon d\nu$  is the integrated area under the fluorophore absorbance spectrum. The radiative rate is directly related to the quantum yield

( $\Phi$ ) and fluorescence lifetime ( $\tau$ ) of the fluorophore ( $k_r = \Phi/\tau$ ). If the quantum yield of the probe is unity, the radiative rate is the inverse of the fluorescence lifetime.<sup>63</sup> Because the radiative rate is related to the square of the refractive index, if the observed change in fluorescence lifetime is due to the changing refractive index with increasing pressure, a plot of  $1/\tau$  vs  $n^2$  should be linear. Figure 22 shows that a Strickler-Berg interpretation of our data reasonably explains ( $r^2 = 0.926$ ) the decrease in fluorescence lifetime of BTBP as the refractive index of the MeOH/CO<sub>2</sub> mixture increases with increasing pressure. Similar results (not shown) have been determined for BTBP in neat CO<sub>2</sub> at 46.0 °C.

Through the Debye-Stokes-Einstein equation, the rotational reorientation time of a solute can be directly related to the physical parameters of the system, including the viscosity, temperature, and the volume of the reorienting species.<sup>56</sup> Figure 23 presents the experimental rotational correlation time for BTBP as a function of cosolvent mixture bulk density at 46.0 °C. The rotational correlation time predicted from the DSE equation (given the volume of BTBP, the temperature, and the calculated bulk viscosity) as well as the rotational correlation time for BTBP in neat CO<sub>2</sub> are denoted. In the low density region, we see that the rotational correlation time deviates significantly from predicted values and only nears the DSE prediction at higher densities. This same phenomena has been observed for BTBP in neat CO<sub>2</sub> at 35 °C<sup>81</sup> and is attributed to a clustering of solvent molecules around the solvent, increasing the fluorophore volume and hence the rotational correlation time. This solute-fluid clustering dissipates at higher liquid-like bulk densities causing the rotational correlation time to more closely follow DSE predictions. The rotational correlation time in our system can allow us to predict the size of the reorienting species as well as determine the solvation of MeOH surrounding the BTBP.

The observed rotational correlation time ( $\phi_{obs}$ ) can be used in conjunction with the rotational correlation time of BTBP in MeOH ( $\phi_{MeOH} = 150$  ps) and the rotational correlation time in neat CO<sub>2</sub> ( $\phi_{CO_2} = 34$  ps) to estimate the fraction of MeOH ( $f_{MeOH}$ ) and CO<sub>2</sub> ( $f_{CO_2}$ ) observed by the probe through the following equation:

$$\phi_{obs} = \phi_{MeOH} f_{MeOH} + \phi_{CO_2} f_{CO_2} \quad (14)$$

Figure 24 presents the percentage of MeOH estimated using this analysis as a function of fluid density. Recall that the bulk MeOH mole fraction is only 5%. We observe a very large degree of preferential solvation (30% MeOH; 6 fold) surrounding the BTBP in the low density region before leveling off in the higher density region. It is interesting to note that the percentage of MeOH in the higher density region is actually less than the expected for a 5 mol% solution. This may in part be due to the oversimplification of Equation 14 in predicting the MeOH solvation.

The degree of “solvent” clustering around the BTBP probe can be roughly estimated by dividing the recovered experimental rotational correlation time ( $\phi_{experimental}$ ) by the rotational correlation time predicted by the DSE equation ( $\phi_{DSE}$ ). Figure 25 presents  $\phi_{experimental} / \phi_{DSE}$  as a function of bulk density for the cosolvent mixture. These average values show that in the low density region, we are observing an increase in the BTBP rotational reorientation on the order of 5 times the expected value. This increase in the rotational correlation time is consistent with local density augmentation of the solvent surrounding the solute. In turn, this is consistent with other systems<sup>81</sup> where clustering phenomena decreases

with increasing bulk density of the mixture.

### Conclusions

For the MeOH-modified CO<sub>2</sub> system we see a significant decrease in the fluorescence lifetime of BTBP with increasing bulk mixture density. This phenomena can be readily explained by use of the Strickler-Berg analysis, showing that the change in BTBP fluorescence lifetime is due to changing solvent refractive index. We observe that the rotational correlation time for BTBP in MeOH-modified CO<sub>2</sub> deviates significantly from the Debye-Stokes-Einstein prediction and the recovered rotational correlation time in neat CO<sub>2</sub>. These deviations can be used to predict the local composition surrounding the BTBP in the MeOH-modified CO<sub>2</sub> system. Preliminary results show that preferential MeOH solvation exists surrounding BTBP in MeOH-modified CO<sub>2</sub> in the low bulk density region and that the local composition surrounding the probe can be estimated at approximately 5 times the bulk density in the low density region.

### **Summary**

Steady-state and time-resolved fluorescence spectroscopy were used to quantify solvation phenomena occurring in several neat and modified supercritical fluid systems. Measurement of pyrene I<sub>1</sub>/I<sub>3</sub> ratios allowed us to experimentally determine solute-fluid interactions occurring in supercritical water. These results showed that near one half the fluid critical density, the local density augmentation is on the order of five times greater than the bulk water density. Pyrene was further used to quantify intermolecular interactions in supercritical *n*-alkanes (fuel precursors). An increase in local alkane density surrounding pyrene dissolved in supercritical C<sub>3</sub>, C<sub>6</sub>, and C<sub>7</sub> *n*-alkanes was observed at one half the critical density with a degree of augmentation on the order of 3 to 4 times the bulk alkane density. These results have shown that further experimentation is needed for a complete picture of how solute-fluid interactions in aviation fuel precursors may affect fuel performance. We have used rotational reorientation measurements of a large neutral solute, BTBP, to track bulk polymer dynamics within several PDMS polymers. We have shown that the addition of ambient temperature CO<sub>2</sub> to PDMS causes a dramatic decrease in the BTBP rotational correlation time near the CO<sub>2</sub> gas-liquid phase boundary (25 °C). In addition, we showed that addition of supercritical CO<sub>2</sub> allowed us to tune the bulk polymer density and the BTBP rotational dynamics. We have also used BTBP rotational reorientation measurements to quantify preferential solvation in MeOH-modified CO<sub>2</sub>. We observed that the rotational correlation time for BTBP in MeOH-modified CO<sub>2</sub> deviated significantly from the Debye-Stokes-Einstein prediction and the recovered BTBP rotational correlation time in neat CO<sub>2</sub>. These deviations were used to predict the local composition surrounding the BTBP in the MeOH-modified CO<sub>2</sub> system. Preliminary results have shown that preferential MeOH solvation exists surrounding BTBP in MeOH-modified CO<sub>2</sub> in the low bulk density region. From these results, the local composition surrounding the probe can be estimated to be approximately 5 times the bulk density in the low density region.

## References

1. U. Van Wassen, I. Swaid, and G.M. Schneider, *Angew. Chem. Int. Ed. Engl.*, **19**, 515 (1980).
2. C.A. Eckert, J.G. Van Alsten, and T. Stoicos, *Environ. Sci. Technol.*, **20**, 319 (1986).
3. D. Bradley, *New Scientist*, **143**, 32 (1994).
4. V.K. Jain, *Environ. Sci. Technol.*, **27**, 807 (1993).
5. S.B. Hawthorne, Y. Yang, and D.J. Miller, *Anal. Chem.*, **66**, 2912 (1994).
6. T.A. Rhodes, K. O'Shea, G. Bennett, K.P. Johnston, and M.A. Fox, *J. Phys. Chem.*, **99**, 9903 (1995).
7. B.C. Wu, M.T. Klein, and S.I. Sandler, *Ind. Eng. Chem. Res.*, **30**, 822 (1991).
8. R. Li, T.D. Thornton, and P.E. Savage, *Environ. Sci. Technol.*, **38**, 321 (1992).
9. T.D. Thornton and P.E. Savage, *AIChE J.*, **38**, 321 (1992).
10. S. Gopalan and P.E. Savage, *J. Phys. Chem.*, **98**, 12646 (1994).
11. P.E. Savage, R. Li, and J.T. Santini, Jr., *J. Supercrit. Fluids*, **7**, 135 (1994).
12. T.J. Houser, D.M.; Tiffany, Z. Li, M.E. McCarville, and M.E. Houghten, *Fuel*, **65**, 827 (1986).
13. P.A. Webley, J.W. Tester, and H.R. Holgate, *Ind. Eng. Chem. Res.*, **30**, 1745 (1991).
14. E.U. Franck, *Ber. Bunsenges. Phys. Chem.*, **88**, 820 (1988).
15. L. Ember, *C&E News*, **72**, 4 (1994).
16. J. Krukowski, *Pollution Eng.*, **26**, 11 (1994).
17. J. Beard, *New Scientist*, **139**, 19 (1993).
18. J. Manji, *Automation*, **38**, 26 (1991).
19. C. Caruana, *Chem. Eng. Prog.*, **91**, 10 (1995).
20. R. Gould, *New Scientist*, **141**, 19 (1994).
21. I.B. Petsche and P.G. Debenedetti, *J. Chem. Phys.*, **91**, 7075 (1989).
22. J.A. O'Brien, T.W. Randolph, C. Carlier, and S. Ganapathy, *AIChE J.*, **39**, 1061 (1993).
23. S. Kim and K.P. Johnston, *AIChE J.*, **33**, 1603 (1987).
24. C.A. Eckert and B.L. Knutson, *Fluid Phase Equil.*, **83**, 93 (1993).
25. K.P. Johnston, S. Kim, and J. Combes, in *Supercritical Fluid Science and Technology*, K.P.L Johnston and J.M.L. Penninger, Eds. (ACS Symposium Series 406, American Chemical Society, Washington, DC, 1989).
26. S.G. Kazarian and M. Poliakoff, *J. Phys. Chem.*, **99**, 8624 (1995).
27. R.W. Shaw, T.B. Brill, A.A. Clifford, C.A. Eckert, and E.U. Franck, *C&E News* **69**, 26 (1991).
28. S. Iyer, G.R. Nicol, and M.T. Klein, *J. Supercrit. Fluids* **9**, 26 (1996).
29. T.J. Houser, Y. Zhou, and X. Liu, *J. Supercrit. Fluids* **9**, 106 (1996).
30. E.U. Franck, *Ber. Bunsenges. Phys. Chem.* **88**, 820 (1988).
31. E.U. Franck, *Pure Appl. Chem.* **53**, 1401 (1981).
32. L.B. Kriksunov and D.B. Macdonald, *J. Electrochem. Soc.* **142**, 4069 (1995).
33. E.E. Brock and P.E. Savage, *AIChE J.*, **41**, 1874 (1995).

34. R. Li, T.D. Thornton, and P.E. Savage, *Environ. Sci. Technol.* **38**, 321 (1992).
35. T.D. Thornton and P.E. Savage, *AIChE J.* **38**, 321 (1992).
36. S. Gopalan and P.E. Savage, *J. Phys. Chem.* **98**, 12646 (1994).
37. P.E. Savage, R. Li, and J.T. Santini, Jr., *J. Supercrit. Fluids* **7**, 135 (1994).
38. T.J. Houser, D.M. Tiffany, Z. Li, M.E. McCarville, and M.E. Houghten, *Fuel* **65**, 827 (1986).
39. P.A. Webley, J.W. Tester, and H.R. Holgate, *Ind. Eng. Chem. Res.* **30**, 1745 (1991).
40. P.T. Cummings, H.D. Cochran, J.M. Simonson, R.E. Mesmer, S. Karaborni, *J. Phys. Chem.* **94**, 5606 (1992).
41. J. Gao, *J. Am. Chem. Soc.* **115**, 6893 (1993).
42. K.P. Johnson, P.B. Balbuena, T. Xiang and P.J. Rossky, in *Innovations in Supercritical Fluids Science and Technology*, K.W. Hutchenson, N.R. Foster, Eds. (ACS Symposium Series 608, American Chemical Society, Washington, DC, 1995).
43. D.C. Dong and M.A. Winnik, *Can. J. Chem.* **62**, 2560 (1984).
44. R. Waris, W.E. Acree, Jr., and K.W. Street, Jr., *Analyst* **113**, 1465 (1988).
45. K. Hara and W.R. Ware, *Chem. Phys.* **51**, 61 (1980).
46. J.K. Rice, E.D. Niemeyer, R.A. Dunbar, and F.V. Bright, *J. Am. Chem. Soc.* **117**, 5832 (1995).
47. Y.-P. Sun, C.E. Bunker, and N.B. Hamilton, *Chem. Phys. Lett.* **210**, 111 (1993).
48. J. Zagrobelny and F.V. Bright, *J. Am. Chem. Soc.* **115**, 701 (1993).
49. Y.-P. Sun and C.E. Bunker, *J. Phys. Chem.* **99**, 13778 (1995).
50. S.-H. Chen and V.L. McGuffin, *Appl. Spectrosc.* **48**, 596 (1994).
51. G. Patonay, M.E. Rollie and I.M. Warner, *Anal. Chem.* **57**, 569 (1985).
52. Schematics and part information for a similar high-pressure titanium cell were kindly provided through a private communication with Professor T.B. Brill and Mr. M. Kieke.
53. For a similar cell design, see: P.D. Spohn and T.B. Brill, *Appl. Spectrosc.* **41**, 1152 (1987).
54. J.M. Harris, F.E. Lytle, and T.C. McCain, *Anal. Chem.* **48**, 2095 (1976).
55. J.N. Demas, *Excited State Lifetime Measurements* (Academic Press, New York, 1983).
56. J.R. Lakowicz, *Principles of Fluorescence Spectroscopy* (Plenum Press, New York, 1983).
57. K.S. Pitzer and S.M. Sterner, *J. Chem. Phys.* **101**, 3111 (1994).
58. P.W. Atkins, *Physical Chemistry*, 4th ed. (W.H. Freeman & Co., New York, 1990).
59. M.J. Kamlet, J.L. Abboud, and R.W. Taft, *J. Am. Chem. Soc.* **99**, 6027 (1977).
60. T.A. Betts, J. Zagrobelny, and F.V. Bright, *J. Am. Chem. Soc.* **114**, 8163 (1992).
61. B. Stevens, M.F. Thomaz, and J. J. Jones, *Chem. Phys.* **46**, 405 (1967).
62. *CRC Handbook of Chemistry and Physics*, 74th Ed. (Chemical Rubber Publishing Co., Boca Raton, 1993-1994).
63. J.B. Birks, *Organic Molecular Photophysics*, Vol. 1 (Wiley Interscience, New York, 1973).
64. B. Stevens and M.F. Thomaz, *Chem. Phys. Lett.* **1**, 535 (1968).

65. J.F. Kauffman, *Anal. Chem.* **68**, 248A (1996).
66. E.D. Niemeyer, R.A. Dunbar, and F.V. Bright, *Appl. Spectrosc.*, submitted for publication.
67. L.H. Sperling, *Introduction to Physical Polymer Science*, 2nd Ed (Wiley Interscience, New York, 1992).
68. P.J. Flory, *Principles of Polymer Chemistry* (Cornell University Press, Ithaca, 1969).
69. Y. Xiong, and E. Kiran, *Polymer*, **36**, 4187 (1995).
70. A.R. Berens, and G.S. Huvarad in *Supercritical Fluid Science and Technology*, K.P. Johnston, J.M.L. Penninger, Eds. (ACS Symposium Series 406, American Chemical Society, Washington, DC, 1989).
71. A. Garg, E. Gulari, and C.W. Manke, *Macromol.*, **27**, 5643 (1994).
72. G.K. Fleming and W.J. Koros, *Macromol.*, **19**, 2285 (1986).
73. J.L. Viovy, C.W. Frank, and L. Monnerie, *Macromol.*, **18**, 2606 (1985).
74. P.D. Hyde, M.D. Ediger, T. Kitano, and K. Ito, *Macromol.*, **22**, 2253 (1989).
75. J.L. Viovy, L. Monnerie, and F. Merola, *Macromol.*, **18**, 1130 (1985).
76. A.D. Stein, D.A. Hoffman, A.H. Marcus, P.B. Leezenberg, C.W. Frank, and M.D. Fayer, *J. Phys. Chem.*, **96**, 5255 (1992).
77. J.P. Jarry and L. Monnerie, *Macromol.*, **12**, 927 (1979).
78. J.P. Jarry, B. Erman, and L. Monnerie, *Macromol.*, **19**, 2750 (1986).
79. P.B. Leezenberg, A.H. Marcus, C.W. Frank, and M.D. Fayer, *J. Phys. Chem.*, **100**, 7647 (1996).
80. A.D. Stein, D.A. Hoffman, C.W. Frank, and M.D. Fayer, *J. Chem. Phys.*, **96**, 3269 (1992).
81. M.P. Heitz and F.V. Bright, *J. Phys. Chem.*, **100**, 6889 (1996).
82. D. Ben-Amotz and J.M. Drake, *J. Chem. Phys.*, **89**, 2 (1988).
83. A.M. Williams and D. Ben-Amotz, *Anal. Chem.*, **64**, 700 (1992).
84. T.A. Betts and F.V. Bright, *Appl. Spectrosc.*, **44**, 1190 (1990).
85. R.D. Spencer and G. Weber, *J. Chem. Phys.*, **52**, 1654 (1970).
86. M.P. Heitz and F.V. Bright, *Appl. Spectrosc.*, **49**, 20 (1995).
87. E. Gratton, D.M. Jameson, and R.D. Hall, *Annu. Rev. Biophys. Bioeng.*, **13**, 105 (1984).
88. D.M. Jameson, E. Gratton, and R.D. Hall, *Appl. Spectrosc. Rev.*, **20**, 55 (1984).
89. F.V. Bright, T.B. Betts, and K.S. Litwiler, *CRC Crit. Rev. Anal. Chem.*, **21**, 389 (1990).
90. J.M. Beecham, E. Gratton, M. Ameloot, J.R. Knutson, and L. Brand in *Topics in Fluorescence Spectroscopy*, Vol. 2, J.R. Lakowicz, Ed (Plenum Press, New York, 1991).
91. F.V. Bright, *Appl. Spectrosc.*, **49**, 14A (1995).
92. R.F.T. Stepto, *Eur. Polym. J.*, **29**, 415 (1993).
93. *Supercritical Fluid Engineering Science. Fundamentals and Applications*, E. Kiran, J.F. Brennecke, Eds. (ACS Symposium Series 514, American Chemical Society, Washington, DC, 1993).
94. *Supercritical Fluid Technology. Theoretical and Applied Approaches in Analytical Chemistry*, F.V. Bright, M.E.P. McNally, Eds. (ACS Symposium Series 488, American Chemical Society, Washington, DC, 1992).

95. M.T. Combs, M. Gandee, M. Ashraf-Khorassani, and L.T. Taylor, *Anal. Chem.*, submitted for publication.
96. K.P. Johnston, K.L. Harrison, M.J. Clarke, S.M. Howdle, M.P. Heitz, F.V. Bright, C. Carlier, and T.W. Randolph, *Science*, **271**, 624 (1996).
97. K.D. Tilly, N.R. Foster, S. Macnaughton, and D.L. Tomasko, *Ind. Eng. Chem. Res.*, **33**, 681 (1994).

## Figure Captions

- Figure 1. Simplified schematic of the high-pressure titanium cell used in this work. Abbreviations: FO, polyimide-coated fiber optic; TF, titanium flange; GW, gold washers; SW, sapphire windows; TB, titanium cell body; and HPF, HiP high-pressure fittings.
- Figure 2. Simplified schematic of the apparatus for performing steady-state fluorescence experiments in SCW. Abbreviations: VP, vacuum pump; FPT, freeze-pump-thaw cell; P, high-pressure syringe pump; PT, pressure transducer; GC, GC oven; HCL, He-Cd laser; L, lens; T, XYZ translator; FO, fiber optic; TC, titanium high-pressure cell; FR, flow restrictor; PC, preheater coil; TH, thermocouple; M, monochromator; D, photomultiplier tube detector; and PC, personal computer.
- Figure 3. Simplified schematic of the apparatus for performing time-resolved fluorescence experiments in SCW. Abbreviations: NL, nitrogen laser; BS, beam splitter; I, iris; L, lens; T, XYZ translator; TC, titanium high-pressure cell; BPF, bandpass filter; PD, photodiode; PMT, photomultiplier tube; OSC, digital sampling oscilloscope; PC, personal computer.
- Figure 4. Normalized emission spectra for pyrene in water at 26.6, 281.6 °C and 2450 psia, and 379.7 °C and 3040 psia.
- Figure 5. Pyrene  $I_1/I_3$  ratios as a function of reduced density for pyrene in water at 26.6, 77.4, 127.4, 203.8, 281.6, 379.7, 384.4 and 398.8 °C.
- Figure 6. Pyrene  $I_1/I_3$  ratio as a function of the dielectric cross term,  $f(\epsilon, n^2)$  for the  $T = 379.7$  °C data. The “ $T_{uncomp}$ ” theoretical line is based on gas and high-density liquid water values. The dashed “ $T_{comp}$ ” line is the theoretical line which has been compensated for the known decrease in  $I_1/I_3$  with temperature.
- Figure 7. Recovered local density augmentation ( $\rho_{local}/\rho_{bulk}$ ) as a function of reduced density for pyrene in SCW at  $T = 379.9, 384.4$ , and  $398.8$  °C.
- Figure 8. Excited-state fluorescence intensity decay traces for pyrene in water at  $T = 29.0, 125.7, 272.2$ , and  $379.5$  °C.
- Figure 9. Density/temperature-dependent pyrene fluorescence lifetimes in water at  $T = 29.0$  (●),  $76.6$  (■),  $125.7$  (▲),  $272.2$  (▼), and  $379.5$  (◆) °C. (Inset) An Arrhenius plot of the lifetime data.
- Figure 10. Pyrene  $I_1/I_3$  ratios as a function of reduced density at  $T_r = 1.01$  in *n*-pentane (Panel A); *n*-hexane (Panel B); and *n*-heptane (Panel C).
- Figure 11. Pyrene  $I_1/I_3$  ratios as a function of the dielectric cross term,  $f(\epsilon, n^2)$ , at  $T_r = 1.01$  in *n*-pentane (Panel A); *n*-hexane (Panel B); and *n*-heptane (Panel C).
- Figure 12. Recovered local density augmentation ( $\rho_{local}/\rho_{bulk}$ ) as a function of reduced density at  $T_r = 1.01$  for pyrene in *n*-pentane (Panel A); *n*-hexane (Panel B); and *n*-heptane (Panel C).
- Figure 13. Recovered rotational correlation time as a function of time for BTBP in PDMS (MW = 28000 g/mol).
- Figure 14. Differential phase angle ( $\Delta$ ) as a function of frequency for BTBP in PDMS (MW = 1250, 3780, and 28000 g/mol) (Panel A); Polarized modulation ratio ( $\Delta$ ) as a function of frequency for BTBP in



PDMS (MW = 1250, 3780, and 28000 g/mol) (Panel B).

- Figure 15. Recovered BTBP rotational correlation times in PDMS as a function of increasing molecular weight.  $M_e$  denotes the entanglement values reported for PDMS in the literature.
- Figure 16. Natural log of the recovered BTBP rotational correlation time as a function of PDMS density. The solid line represents the first order fit to the data.
- Figure 17. Recovered BTBP rotational correlation times in PDMS (MW = 1250, 9430, 13650, and 28000 g/mol) as a function of CO<sub>2</sub> pressure at 25 °C.
- Figure 18. Calculated densities of CO<sub>2</sub>-diluted PDMS (MW = 1250, 9430, 13650, and 28000 g/mol) as a function of CO<sub>2</sub> pressure at 25 °C.
- Figure 19. Recovered BTBP rotational correlation times in PDMS (MW = 9430 g/mol) as a function of CO<sub>2</sub> pressure at ambient (25 °C) and supercritical (T = 36.5 °C) conditions (Panel A); Expanded view of the T = 36.5 °C data.
- Figure 20. Steady-state emission and excitation spectra for BTBP in MeOH-modified CO<sub>2</sub> at 46.0 °C and 2667 psia.
- Figure 21. BTBP fluorescence lifetime as a function of cosolvent mixture bulk density at 46.0 °C.
- Figure 22. Strickler-Berg plot for BTBP in MeOH-modified CO<sub>2</sub> at 46.0 °C.
- Figure 23. Recovered BTBP rotational correlation times in MeOH-modified CO<sub>2</sub> at 46.0 °C. The dashed line indicates the rotational correlation times predicted by the Debye-Stokes-Einstein equation. The recovered rotational correlation time in neat CO<sub>2</sub> at the experimental temperature is marked for reference.
- Figure 24. Percentage of MeOH surrounding BTBP calculated using Equation 15 as a function of bulk solvent mixture density.
- Figure 25. Recovered local density augmentation ( $\phi_{\text{experimental}}/\phi_{\text{DSE}}$ ) for BTBP in MeOH-modified CO<sub>2</sub> as a function of bulk cosolvent mixture density.

The diagram illustrates the experimental setup for studying the spin Hall effect of light. At the top, a laser source (FO) and a target (TF) are positioned. A beam splitter (SW) and a waveguide (GW) are used to direct the light. The beam path is indicated by arrows. At the bottom, a detector (HPF) and a target (TB) are positioned. The setup is designed to measure the spin Hall effect of light by observing the deflection of the beam path.

[illegible]

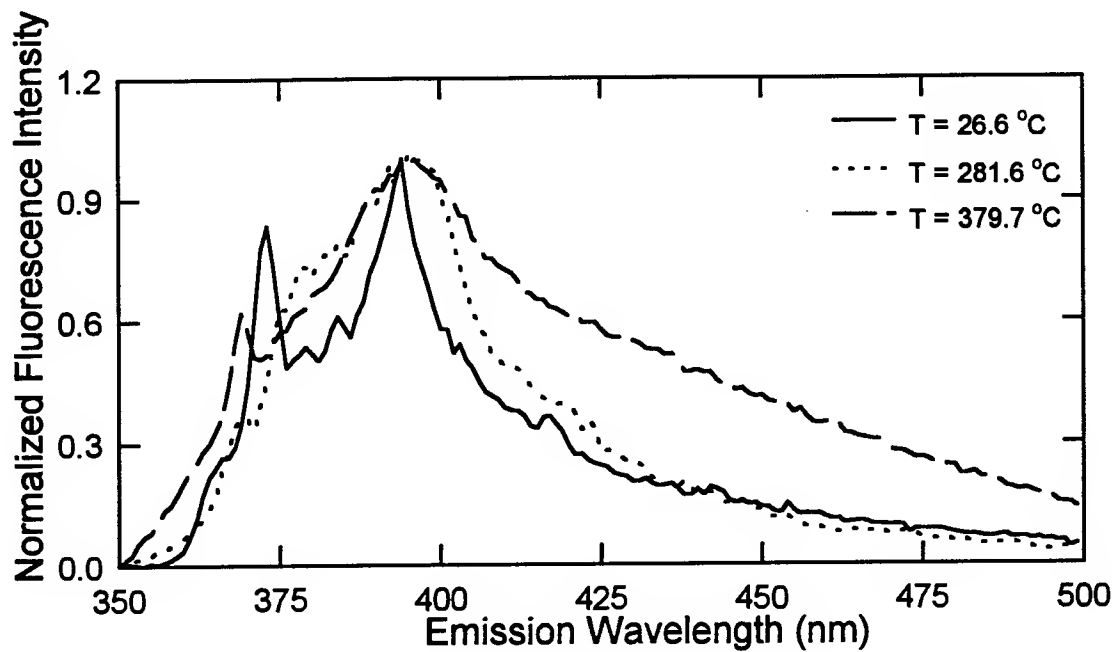


Figure 4

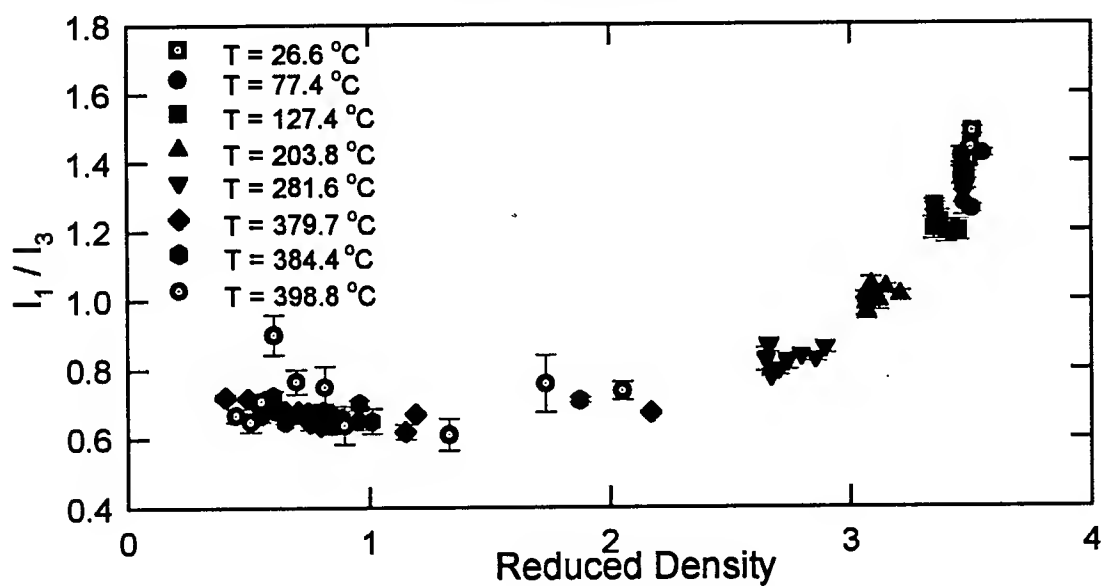


Figure 5

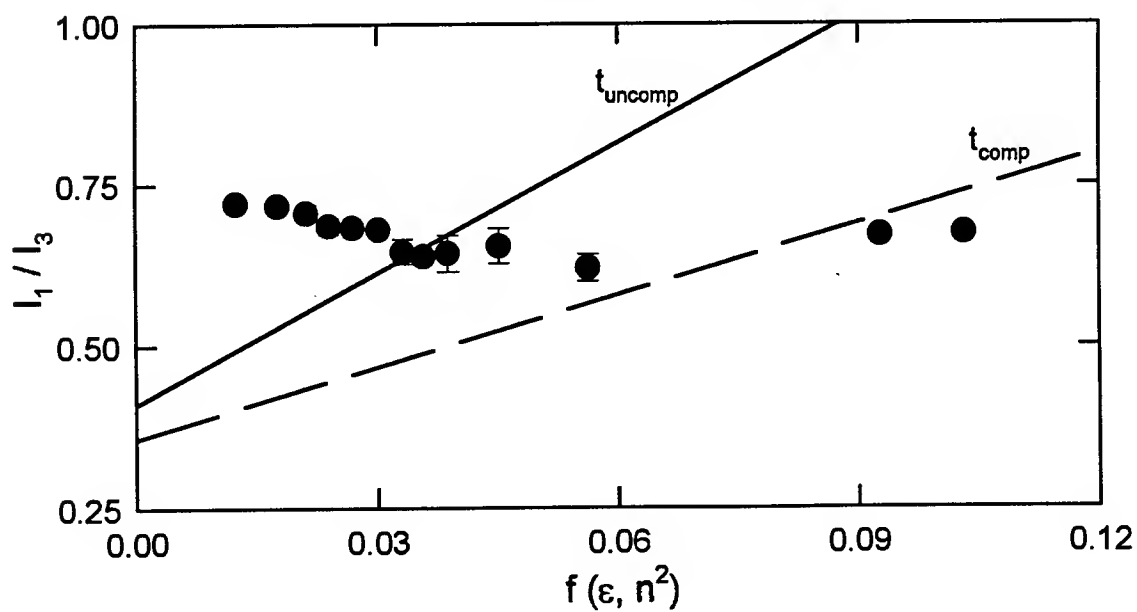


Figure 6

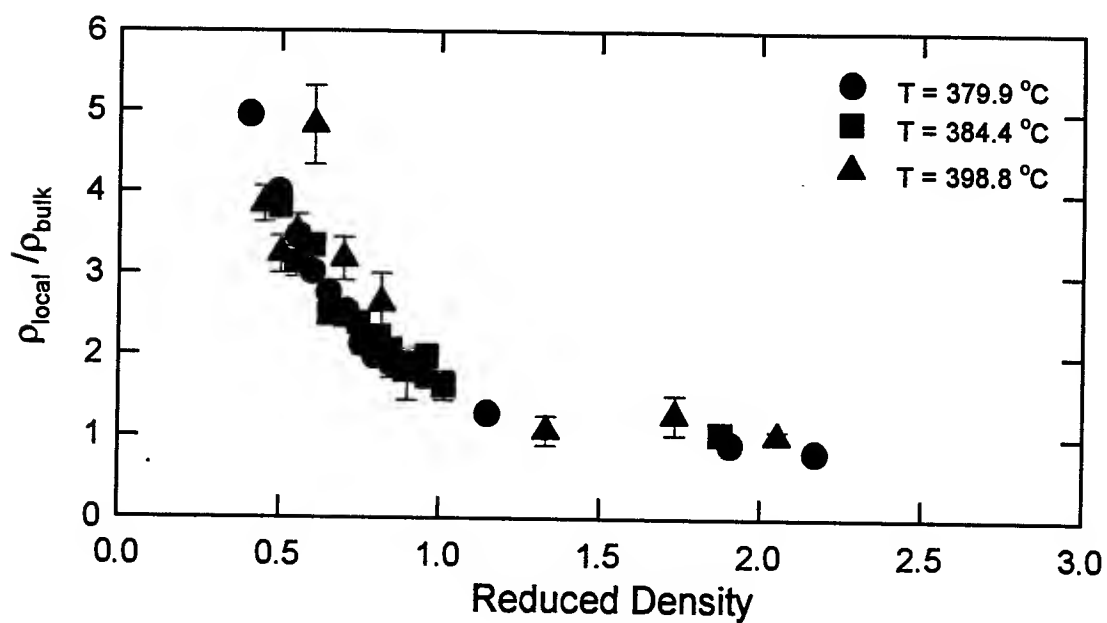


Figure 7

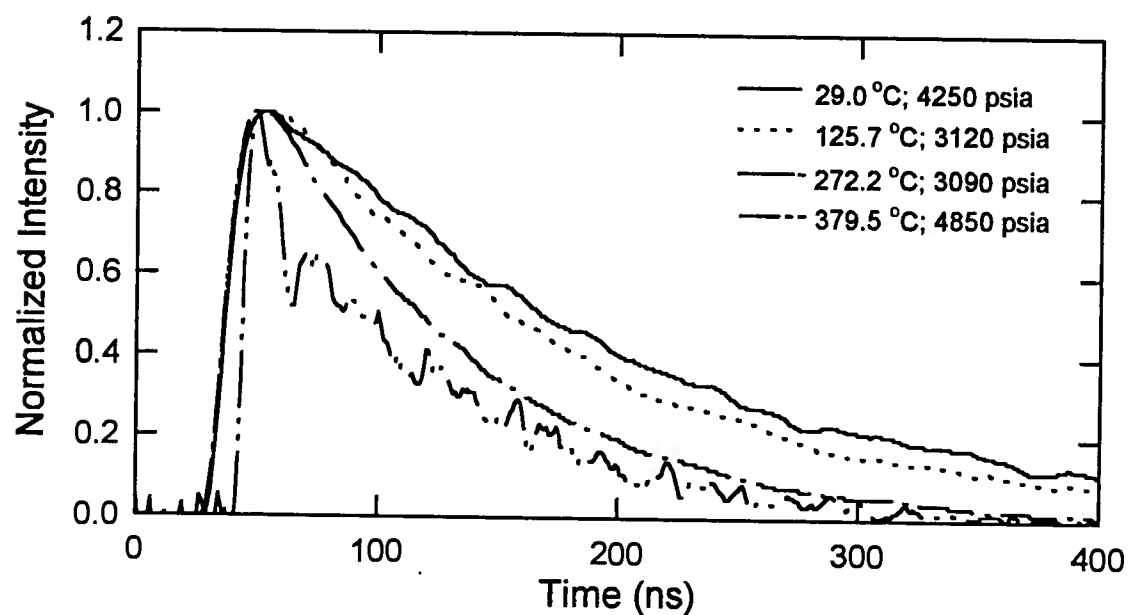


Figure 8

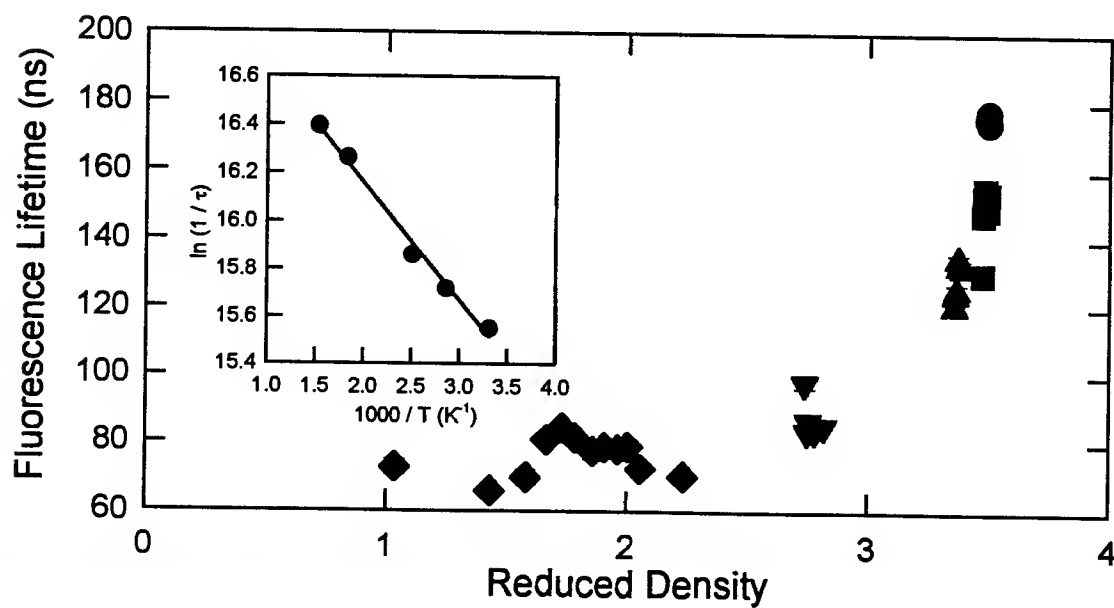


Figure 9

**Figure 10**

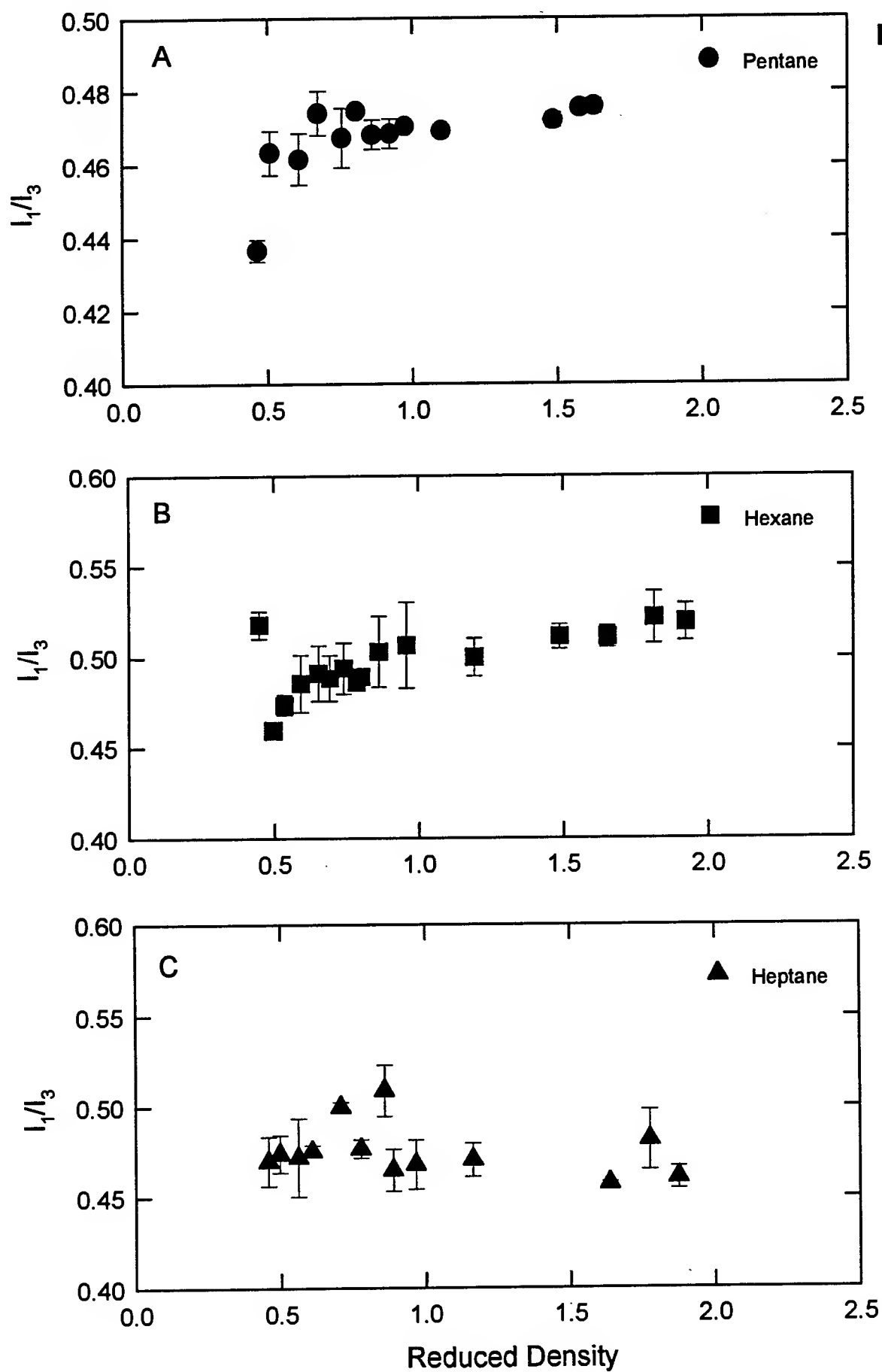


Figure 11

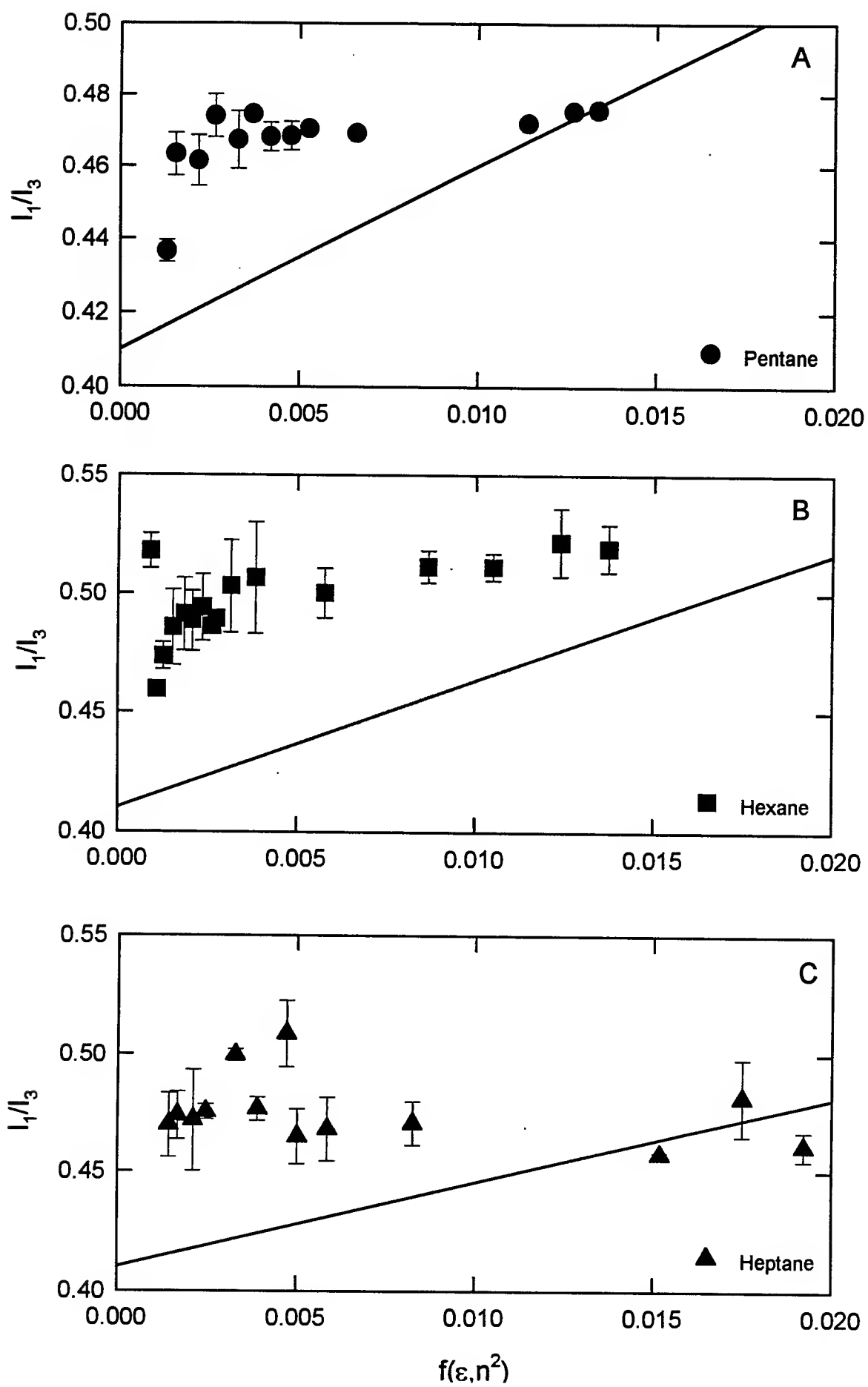
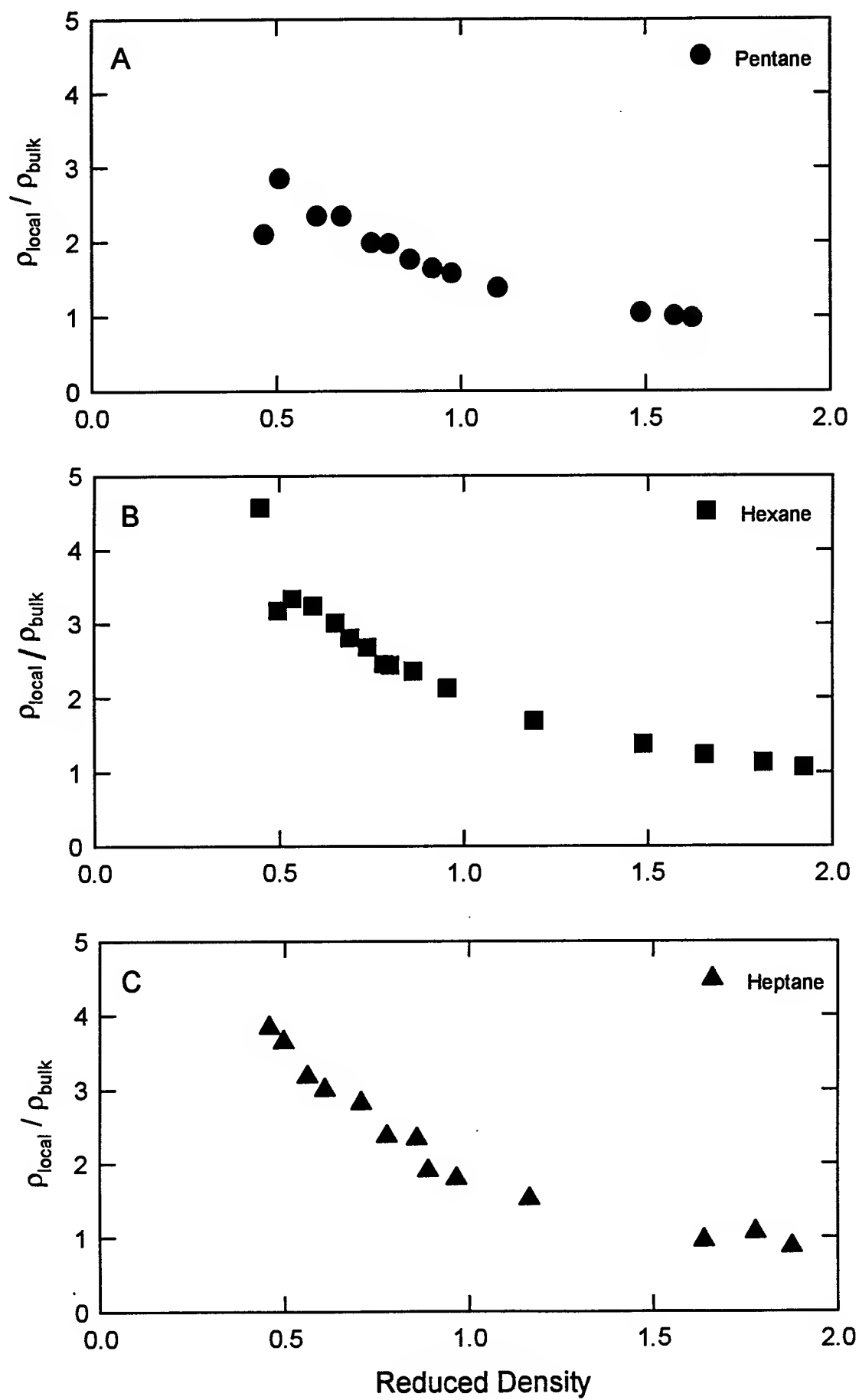


Figure 12



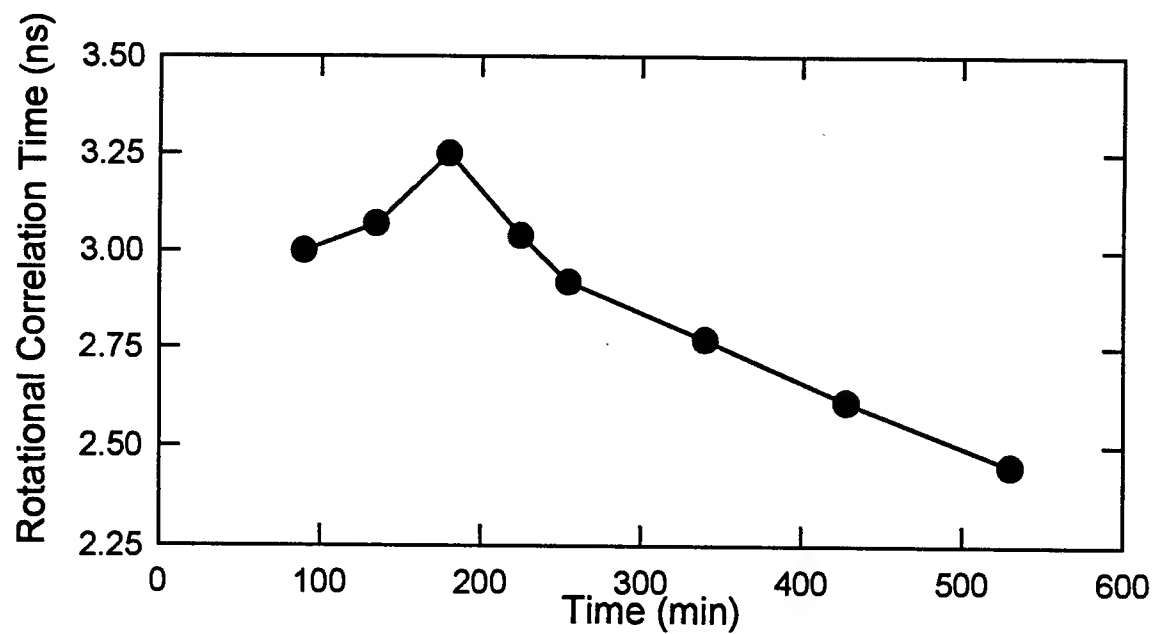


Figure 13

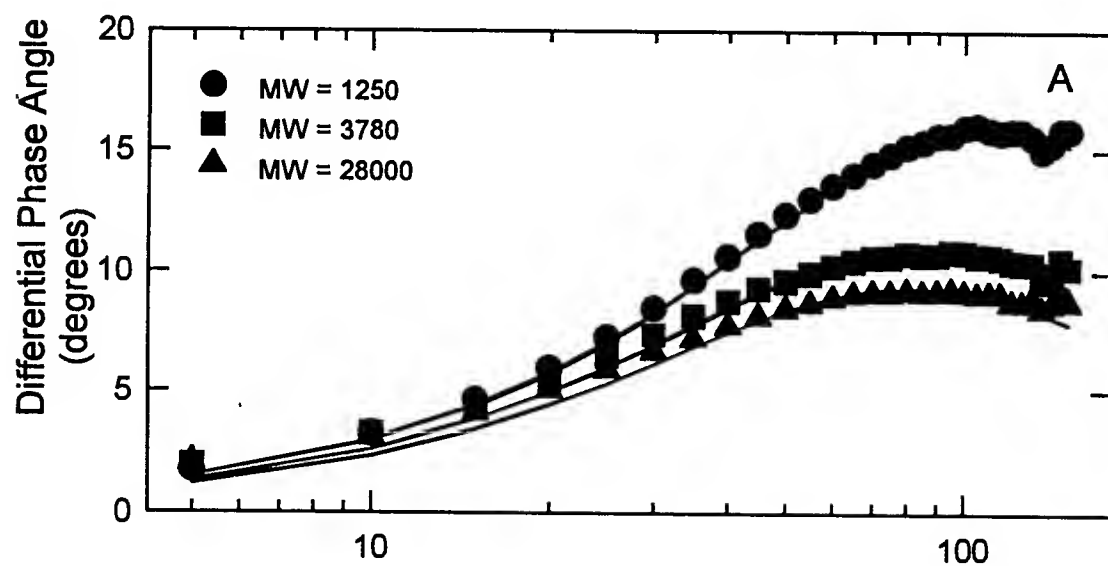
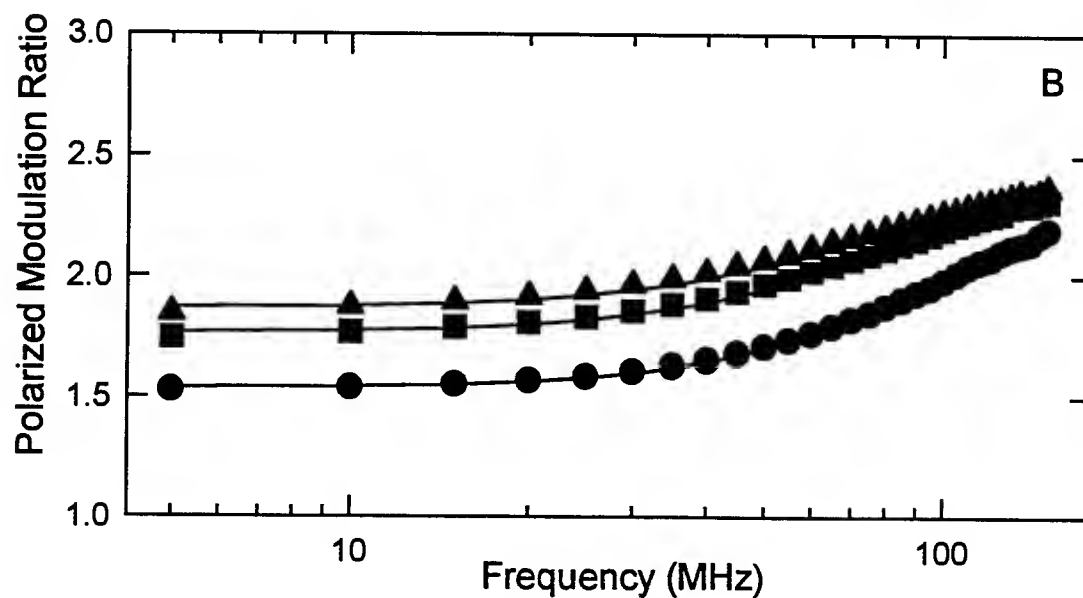


Figure 14





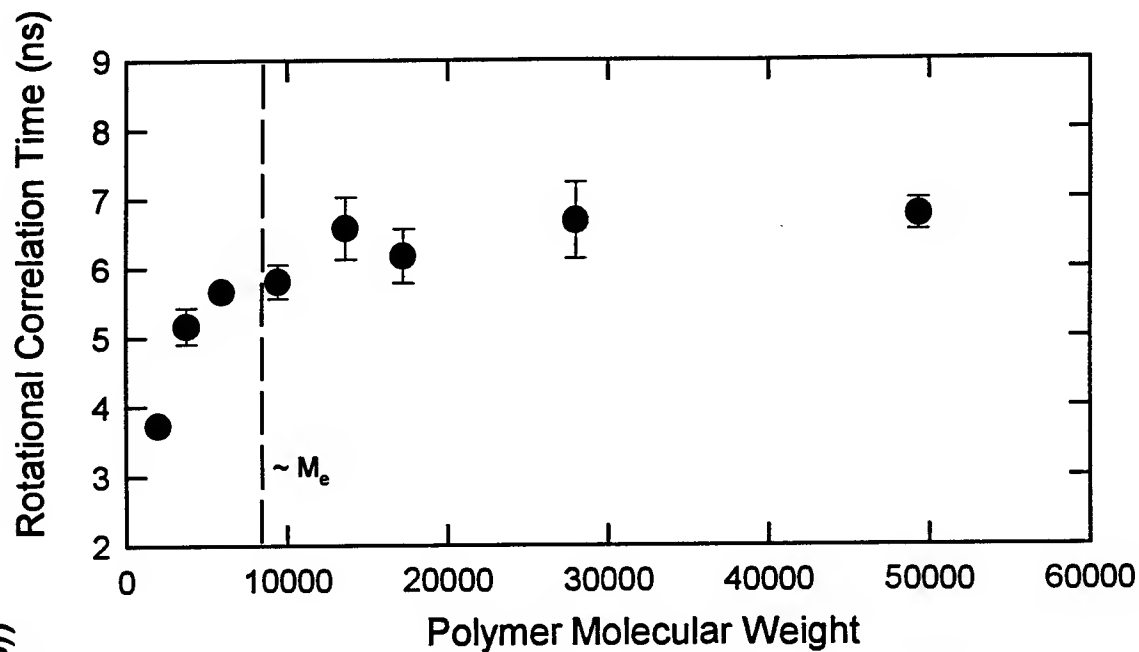


Figure 15

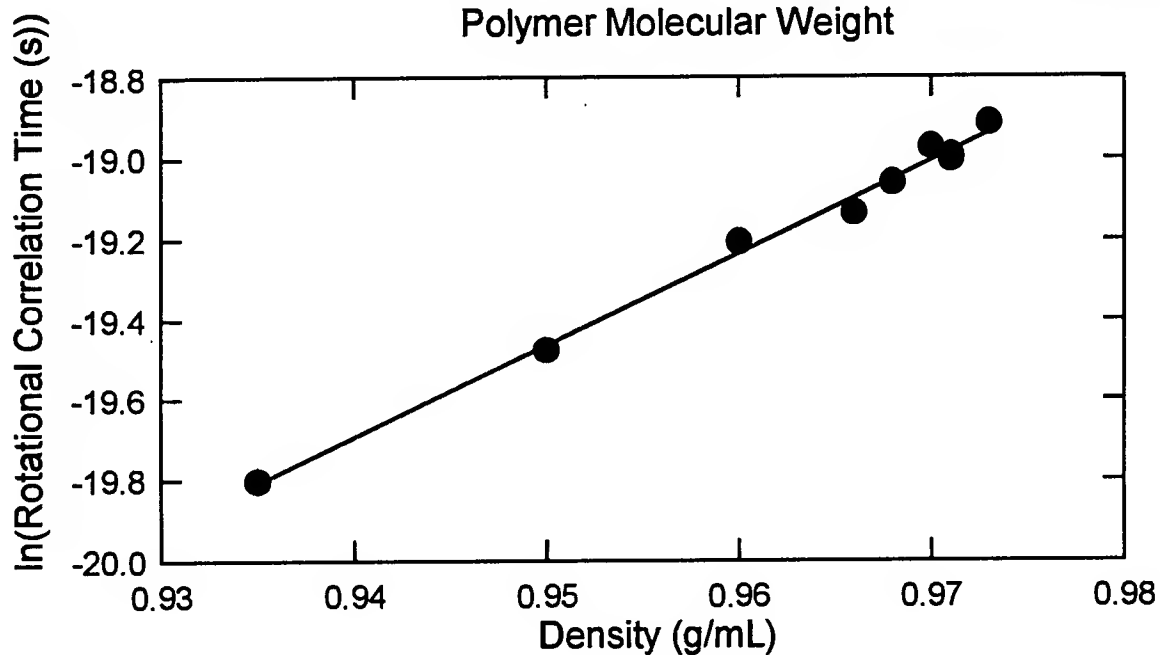


Figure 16

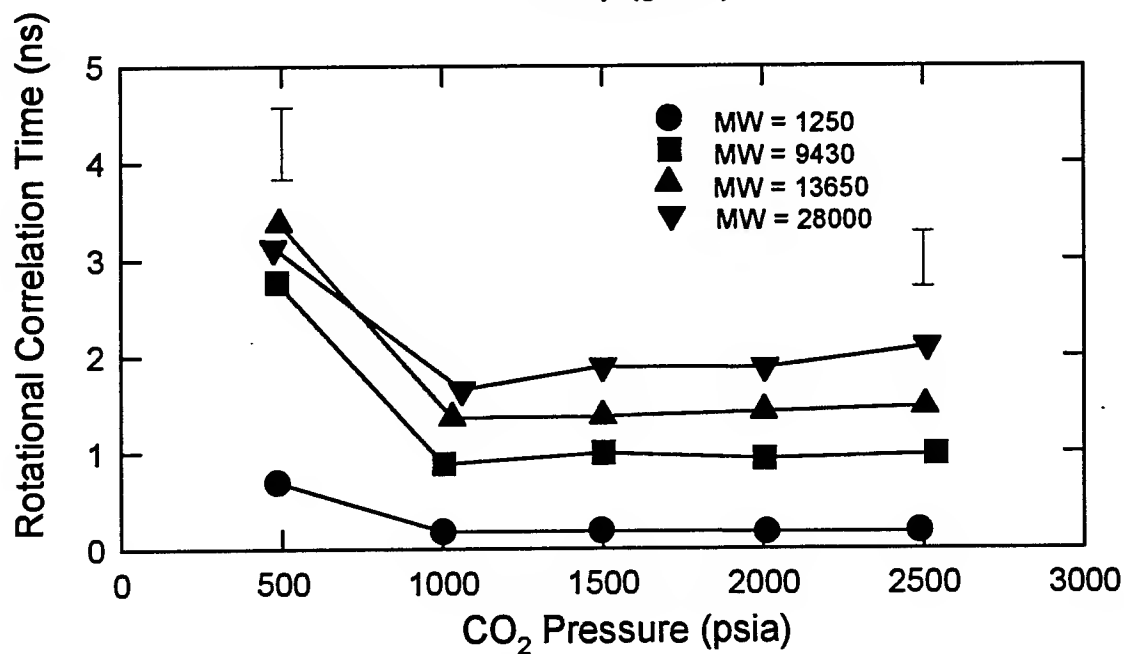


Figure 17

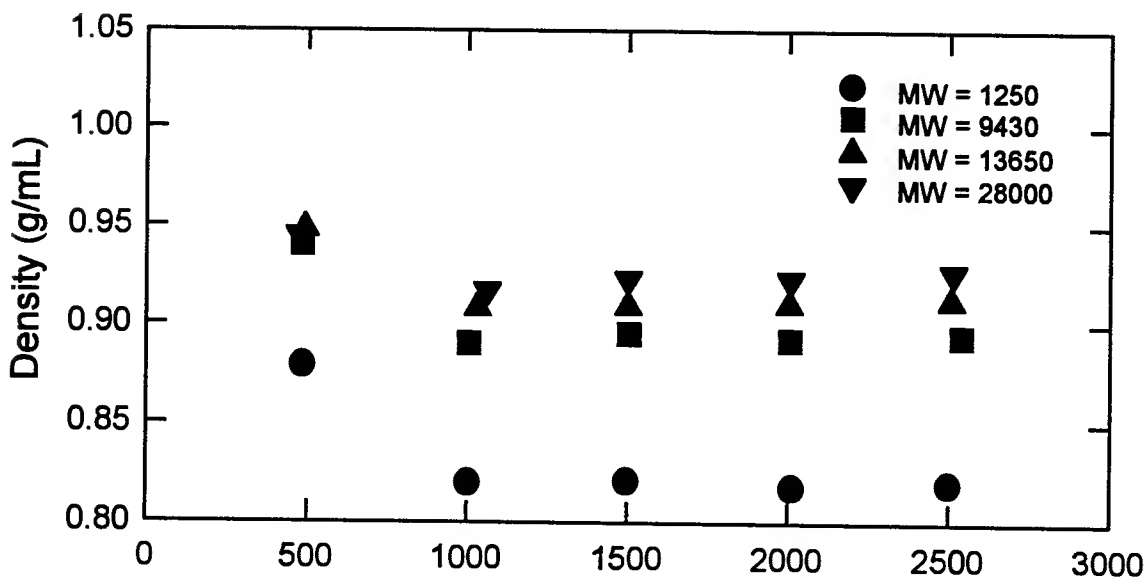


Figure 18

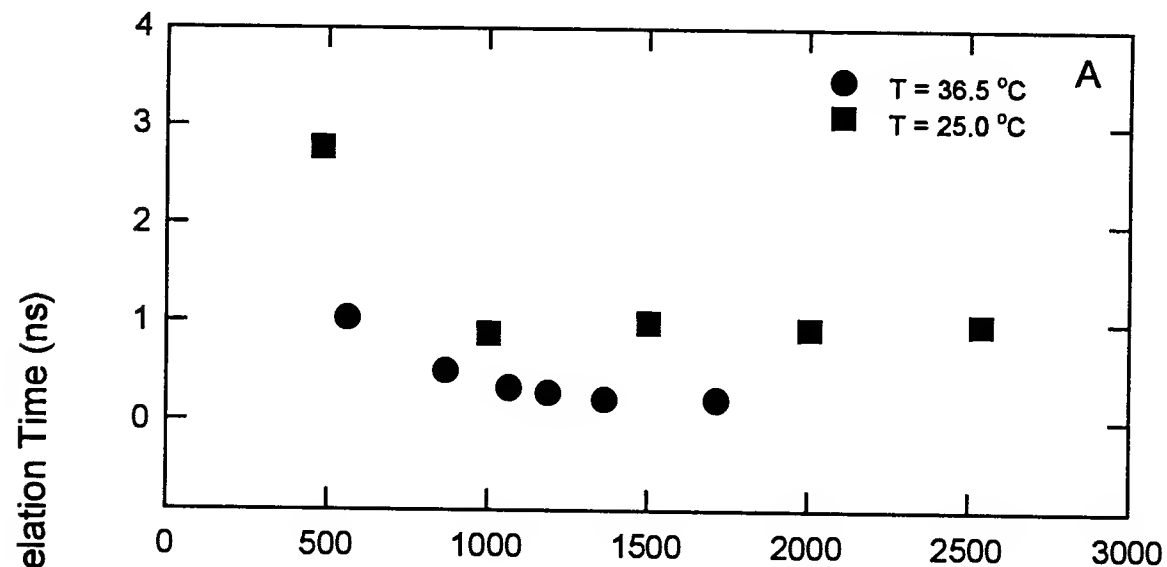
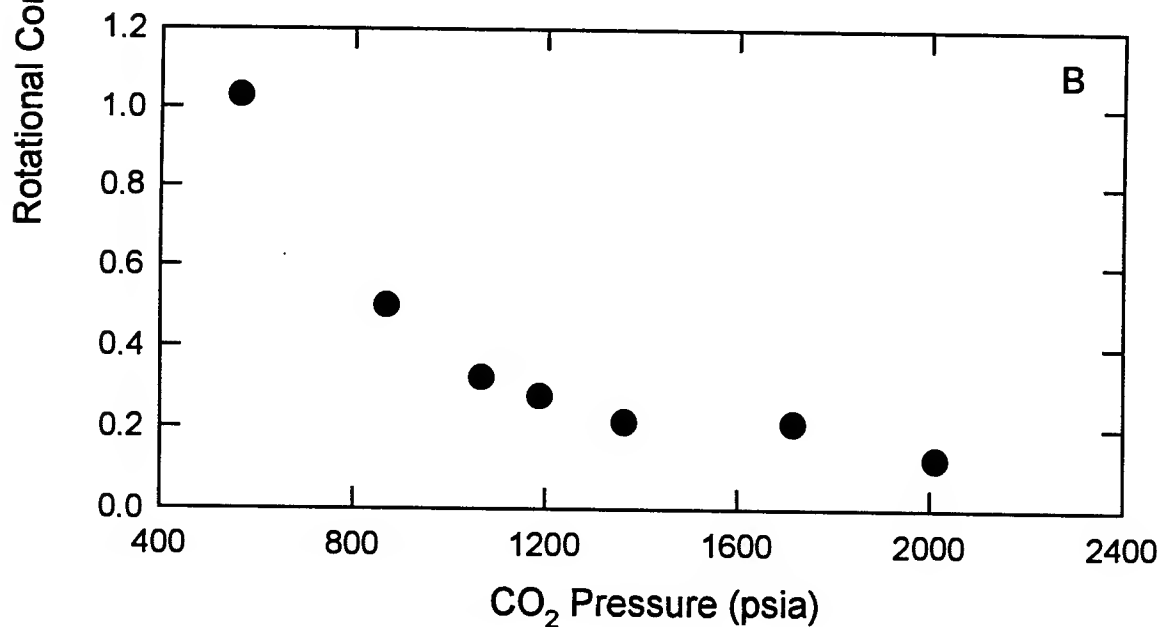
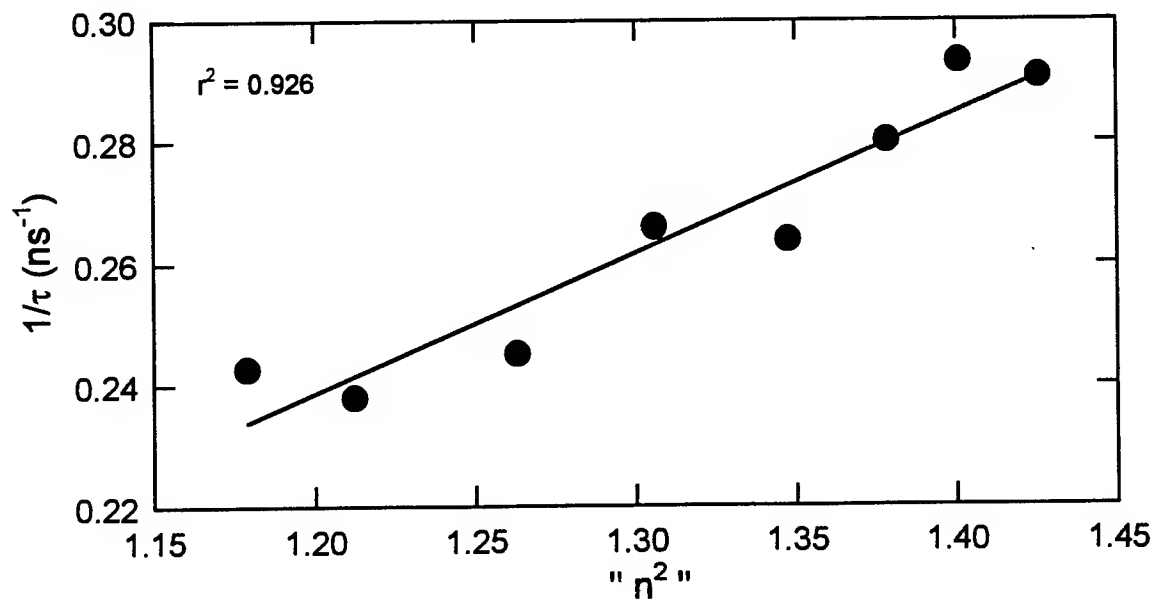
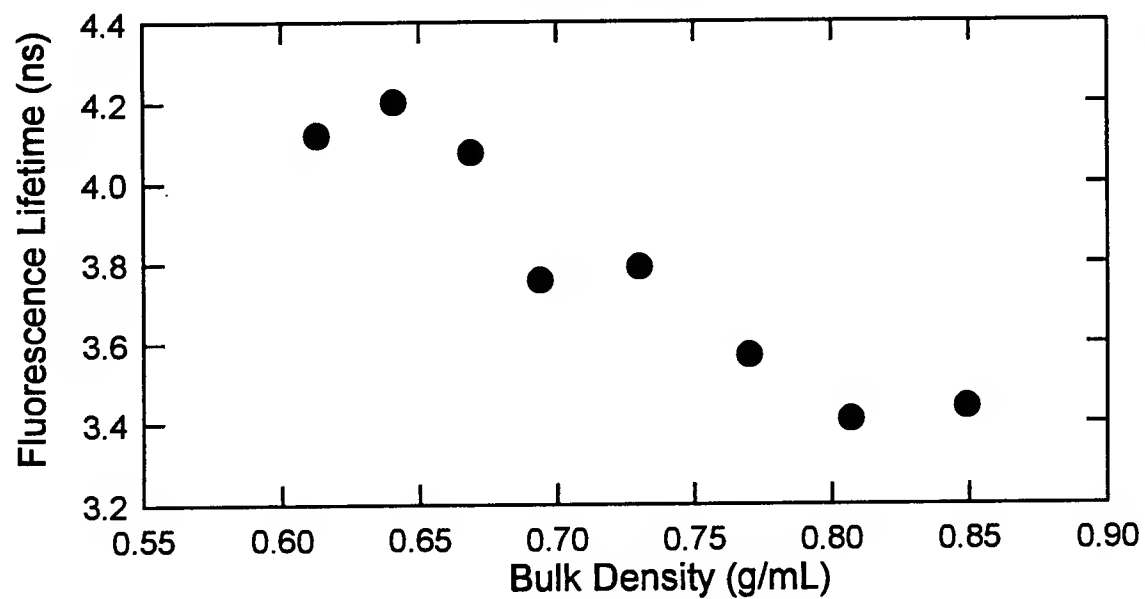
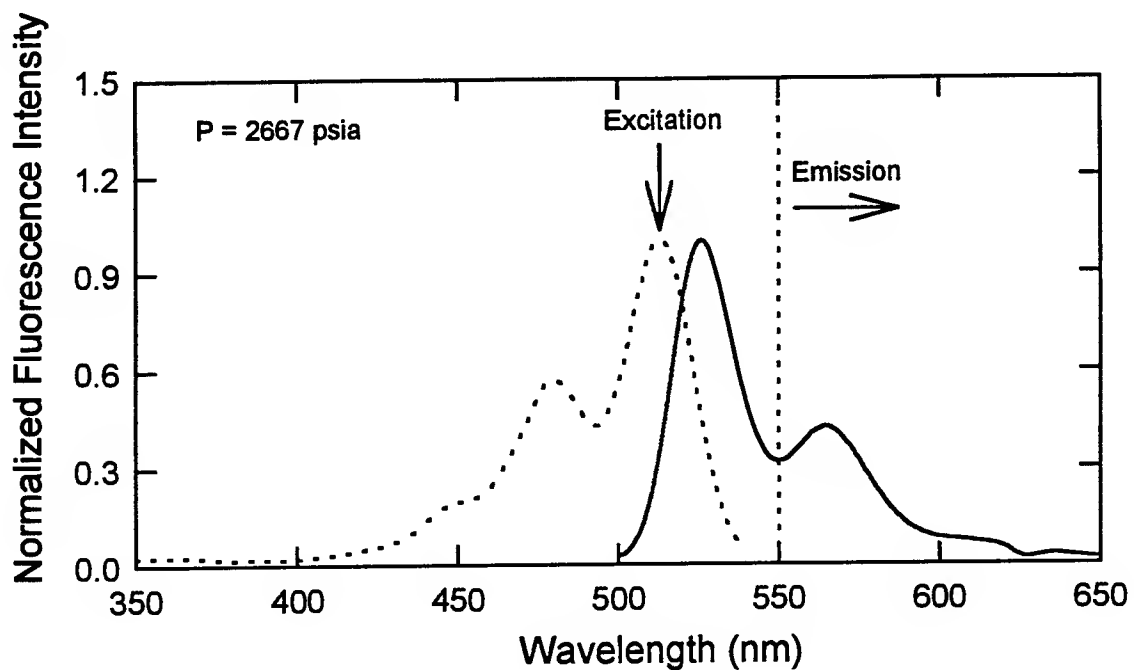


Figure 19





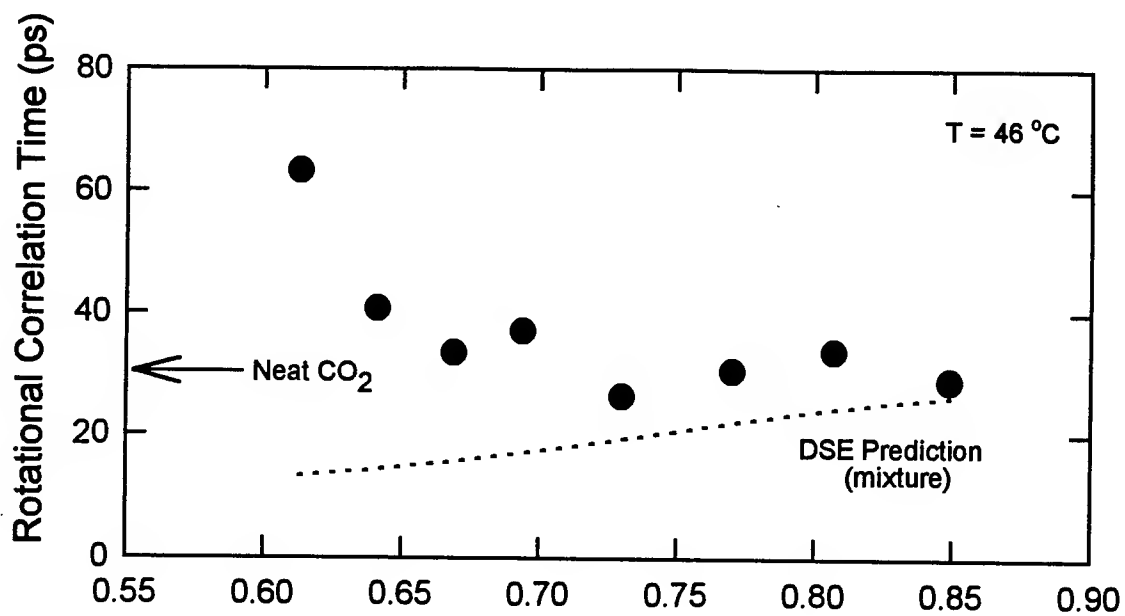


Figure 23

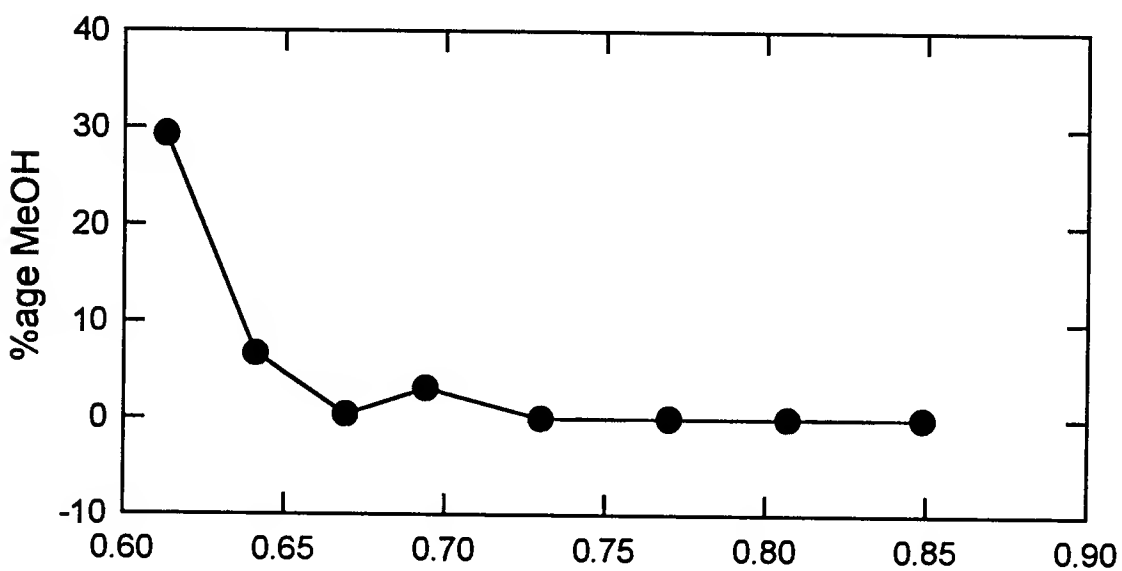


Figure 24

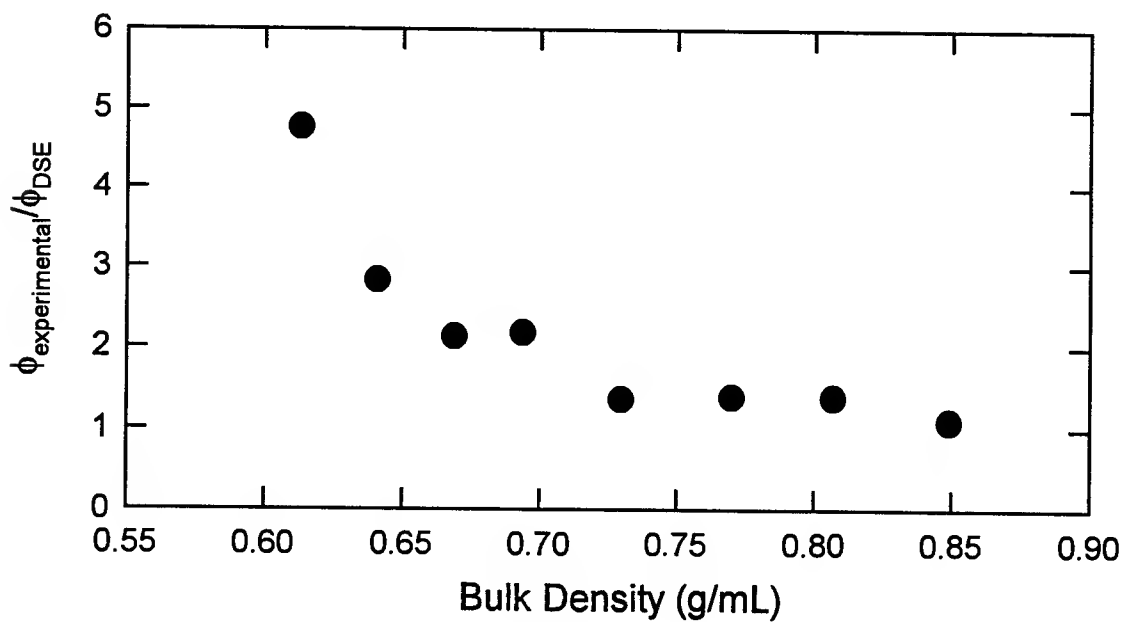


Figure 25

# IMPACT INITIATION OF TRITONAL AND PBX N109

Keith M. Roessig  
Graduate Student  
Department of Aerospace & Mechanical Engineering

University of Notre Dame  
356 Fitzpatrick Hall  
Notre Dame, IN 46556

Final Report for:  
Summer Research Extension Program

Sponsored by:  
Air Force Office of Scientific Research  
Bolling Air Force Base, DC

November 1994

# IMPACT INITIATION OF TRITONAL AND PBX N109

Keith M. Roessig  
Graduate Student  
Department of Aerospace & Mechanical Engineering  
University of Notre Dame

## Abstract

Recently, the behavior of explosive materials under intermediate loading rates ( $10^0$ - $10^4$  s<sup>-1</sup>) has become of interest. Shock and quasi-static behaviors have been well characterized, but in weapons such as a deep earth penetrator, loading conditions in intermediate strain ranges will occur previous to initiation and affect the response of the explosive. This paper will examine the mechanical behavior of a cure cast simulant, Filler-E, and a melt cast simulant under three widely different strain rates. These tests will include quasi-static loading rates on an MTS machine, low velocity tests on a mechanical press, and high velocity tests incorporating a Kolsky bar apparatus at the University of Notre Dame. Impact tests of the explosives Tritonal and PBX-N109 conducted at the Advanced Warhead Experimentation Facility at Eglin AFB, FL will be discussed.

Filler-E and Tritonal are shown to be a very brittle material, while the cure cast simulant and PBX N109 are much more ductile. All these materials have strengths well below that of steel, usually 4340 is used in bomb casings, and can be neglected in comparison. Slight strain rate hardening is seen in both simulants, though for Filler-E, failure strains drop tremendously as failure becomes a dynamic fracture event. The impact tests on Tritonal and PBX-N109 show that these materials need to have high hydrostatic loads after failure to generate the internal friction needed for initiation.

# IMPACT INITIATION OF TRITONAL AND PBX N109

Keith M. Roessig

## 1 Introduction

Explosives are used in a wide variety of applications, from mining to military weaponry to metal forming. Understanding their behavior is very important in all applications to allow for their efficient and safe use. Safety is especially critical in the handling and storage of explosives, and preventing sympathetic, or unplanned, detonation is one of the primary goals of the explosive engineer. In addition, design applications also require a thorough knowledge of the mechanical and thermal behaviors of these materials.

The theory of detonation of explosives is a very complicated one, and, due to the small time scales involved, has not been studied extensively by many academic institutions. A key element of detonation theory is the coupling between mechanics and chemistry during the reaction. This coupling cannot be ignored and should be included in any realistic model [6]. Frequently, the ignition is assumed to be heavily dependent on the formation of hot spots, localized regions of intense heat generation. Factors that can lead to hot spot formation and ignition include jetting, void collapse, viscous heating, shock interaction, internal friction and adiabatic shear localization [2, 6]. Adiabatic shear localization occurs at high strain rates when shear deformation may cause thermal softening through the plastic work done on the material. The softer material deforms more, causing further heating. This self feeding process can become localized into a very small region, and the local temperatures can become very high. Shear banding has been relatively unexplored experimentally as a source of ignition in solid explosives. Field et al. [10] took high speed photographs of explosives initiating from hot spot formation produced by the variety of mechanisms listed above. Evidence of shear localization was found in some tests. Boyle et al. [3] investigated ignition of certain explosives under combined pressure and shear conditions and found that shear bands do form in the interior of the explosive when hydrostatic pressure is applied. Chou [5] has run numerical simulations of the impact of various explosives with steel projectiles and found shear bands to form. Temperatures within these bands, however, were not always great enough to cause initiation. Understanding how these mechanisms interact under high strain rates is essential to the proper design and safe use of reactive materials.

Recently, greater emphasis has been placed on determining the mechanical and reactive properties of explosives deforming under lower strain rates. Initiation under very high strain rate loadings such as shock waves occur due to adiabatic, compressive heating, while initiation under very small, or quasi-static, loading rates occurs mainly from external heating. Loading cases more typically seen by these materials result in strain rates between these extremes during use. One example is the deep earth penetrator. The explosives

inside the weapon will undergo various loading rates and loading geometries before reaction occurs. How the explosives behave during the loading and at detonation by the fuse is critical to successful implementation of these weapons.

The explosives Tritonal and PBX-N109 are examined in this paper. These are respectively a melt cast trinitrotoluene (TNT) based explosive and a cure cast plastic bonded explosive (PBX) containing the Royal Development Explosive (RDX) trinitro triazacyclohexane. No pressed PBX explosives were used. In the interest of safety, the inert melt cast simulant, Filler-E, and inert cure cast PBX simulants were used in the punch tests at the University of Notre Dame. It is important to study these materials for several reasons: 1) to determine the behavior of these materials, 2) determine how closely they imitate the mechanical behavior of the explosives they simulate, 3) gain knowledge of the mechanical behavior without reaction for safer experiments, and 4) record data to be used in finite element codes such as EPIC or ABAQUS to model other experiments where they are used.

Though shear localization can be described by a material process, geometry plays an important role in the event. Failure can be constrained to a shear dominated mode by the experimental setup. Punching and plugging are two situations in which adiabatic shear localization can occur. Plugging is the process in terminal ballistics in which failure occurs in a shear dominated mode, though bending deformation can be quite large. The clearance between the projectile and any support is very large, so in effect it is an infinite plate. Usually plugging occurs at large impact velocities, depending upon the material of the target and projectile. Punching, on the other hand, requires a very small clearance between punch and die, and usually occurs in manufacturing. Punching velocities are small and the failure is again largely dominated by shear since the die prevents much bending deformation. Figure 1 shows the geometric difference between the punching and plugging. The punch test was chosen here because higher strain rates can be obtained at lower punch velocities with the shear deformation constrained to a smaller area.

Load-displacement diagrams are very important in the analysis of punch tests on materials. The various

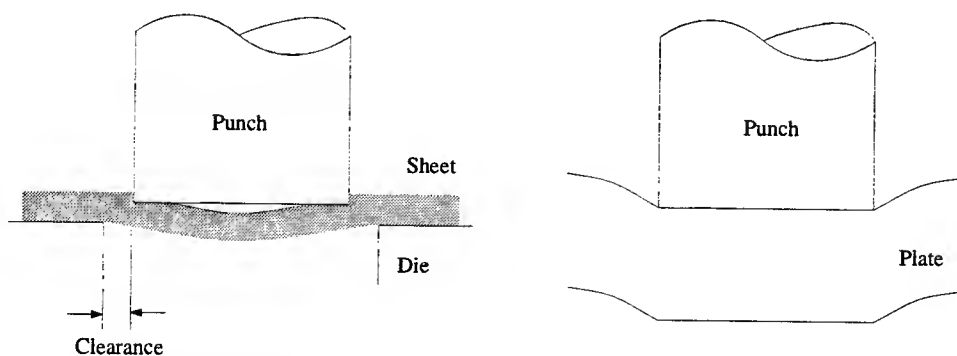


Figure 1: Geometries of punching and plugging operations



sections of the load displacement diagram indicate what may be occurring during each phase of the failure process, see Figure 2. The energy used in the process, the area under the curve, can be correlated to the constitutive characteristics of the material. Yield and ultimate strengths as well as strain hardening values can be estimated from these curves. Most of the area under the graph is energy being used to deform the material, both elastically and plastically. Once, the material begins to fracture, however, most of the required punching energy comes from friction. This is a schematic, and actual load-displacement curves will vary depending on punch velocity and material. Average stress-average strain curves can be determined from the load-displacement data.

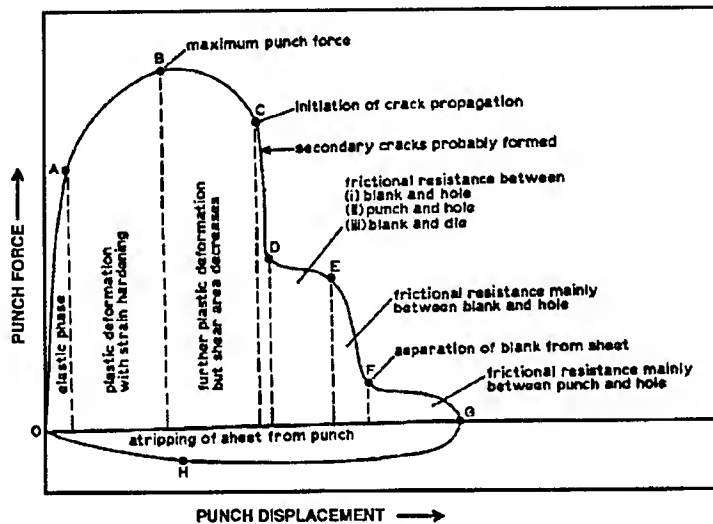


Figure 2: Load-displacement curve schematic (Bai and Johnson, 1982)

## 2 Experimental Method

### 2.1 Tests on Simulants

Three punch tests were conducted at different velocities of  $2.0 \times 10^{-5}$  m/s, 1 m/s, and  $\sim 10$  m/s using the insert/die configuration shown in Figure 3. The insert slides into the die, and the simulant is placed in a recess machined for the specimen. The projectile slides through the hole in the cover plate, which is bolted to the die, and can easily pass through the insert and die.

The melt cast simulant, Filler-E, and cure cast PBX simulant were used in this study. All the specimens were discs of the same size, 0.25" thick and 2" in diameter. These materials were obtained from the High Explosive Research and Development (HERD) division of the Wright Labs Armament Directorate at Eglin AFB. The clearance between the punch and die in all the tests was 2.54mm. Three separate testing methods were used to determine load-displacement curves at a wide range of strain rates.

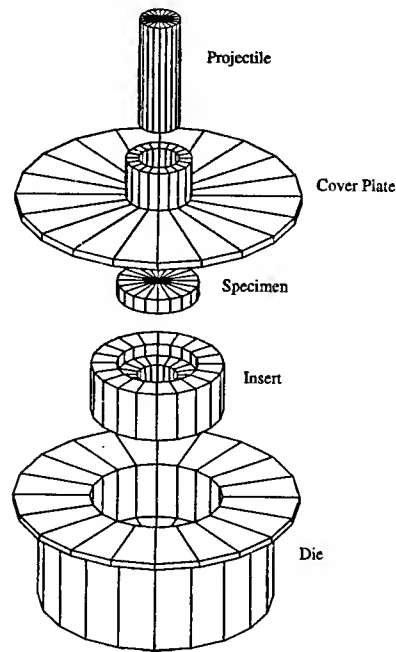


Figure 3: Insert/die configuration for MTS and mechanical press

Quasi-static tests were conducted on a MTS 810 20 kip tension/compression machine at an average shear strain rate of  $5.0 \times 10^{-3}$ . The load was read directly from the load cell on the MTS, while the displacement was obtained through a Epsilon Technology Corp. 3540-1000-ST deflectometer. This instrument is similar to an extensometer, but uses strain gages to measure the deflection of a small shaft. Using the calibration from the manufacturer, voltage traces can be recorded on a digital oscilloscope and then converted to displacements. Though the MTS machine does have its own position indicator, it was found that there was too much error from the machine compression of the projectile and die to measure such small displacements. Voltages were recorded on digital oscilloscopes and then transferred to a PC for analysis.

The low velocity tests were conducted on a mechanical press machine shown in Figure 4. The same die configuration was used on the mechanical press as the MTS machine. The punch velocity was approximately 1 m/s, giving an average strain of  $\sim 200$ . Strain gages were placed on the punch to determine the load upon the specimen during the test. Output traces were recorded on a digital oscilloscope. The displacement was recorded with the same deflectometer used with the MTS machine.

For the high speed tests, a punch-loading Kolsky bar apparatus was built to perform the punch tests. This apparatus is described in volume 8 of the the ASM Handbook [1] and had been successfully used by Dowling et al. [8] and Zurek [13]. The basic design is shown in Figure 5.

Using an air gun a long projectile is fired at the Kolsky bar. This sends a stress pulse down the length

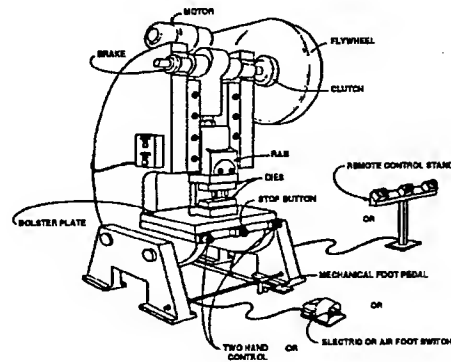


Figure 4: Mechanical press used for low velocity tests

of the bar that can be measured using strain gages placed on the bar as shown. As the pulse reaches the end of the bar, the bar will impact the specimen. The stress pulse will be partially reflected and partially transmitted. The reflected portion of the pulse is measured by the strain gages that also measured the initial stress pulse. While the transmitted portion could be measured using strain gages on a die tube behind the specimen, it is not needed in this case; due to the large diameter of the die as compared to the bar, less than 1% of the pulse is transmitted to the die. It is consequently assumed that there is no transmitted wave, thus simplifying the analysis of the results (section 2.1.1). Average strain rates of up to  $1.0 \times 10^4$  can be obtained with the apparatus at Notre Dame.

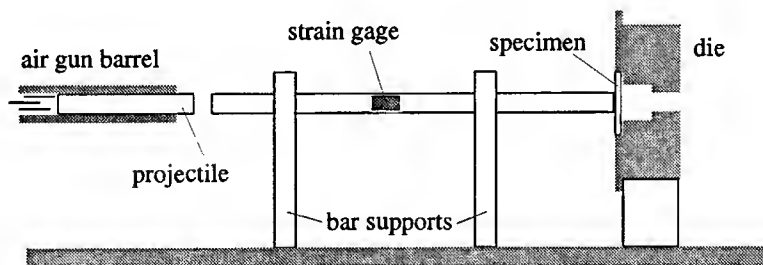


Figure 5: Punch-loading Kolsky bar apparatus

### 2.1.1 Elastodynamic Analysis

In the Kolsky bar experiment, the following analysis [13] may be used to get load-displacement and average stress-average strain curves. During use of a Kolsky bar, Figure 6, stress pulses are sent down the length of the bar to cause a specimen to fail. From elastodynamics, the stress in an elastic wave can be related to the

particle velocity by the relation

$$\sigma = E \frac{\partial u}{\partial x} = \frac{E}{c} \frac{\partial u}{\partial t} = \rho c \frac{\partial u}{\partial t}$$

where  $E$  is the elastic modulus,  $c$  is the elastic wave speed,  $\sigma$  is the axial stress, and the term  $\partial u / \partial t$  describes the particle velocity,  $V$ , in the bar. This velocity will have a different sign depending on the direction of travel (sign of  $c$ ) of the pulse itself and the sign of the stress, and there are two relations describing the stress in relation to the particle velocity.

$$\sigma = \begin{cases} \rho c V, & \text{left traveling wave} \\ -\rho c V, & \text{right traveling wave} \end{cases}$$

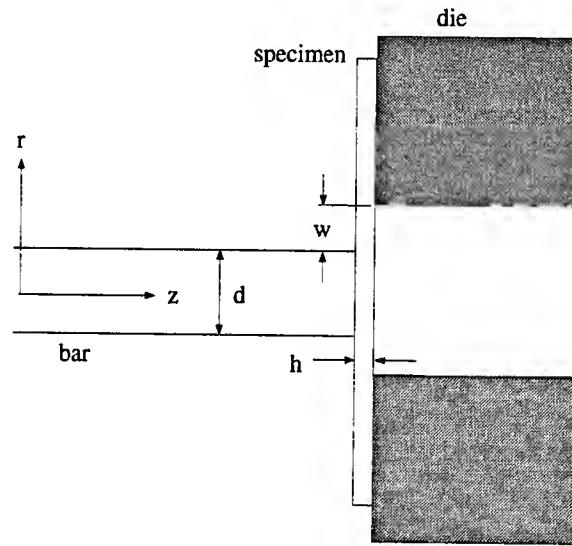


Figure 6: Punch-loading Kolsky bar apparatus

For the incident pulse on the Hopkinson bar, the right traveling wave relations are used, so the incident stress is

$$\sigma_i = -\rho c V_i.$$

Because the bar remains elastic, the strain can be given by

$$\epsilon_i = \frac{\sigma_i}{E}$$

and then substituted into the stress equation to yield

$$V_i = -\frac{E \epsilon_i}{\rho c}$$

The relations for the reflected wave are similar, but these are left traveling waves, so the elastic stress is defined as

$$\sigma_r = \rho c V_r$$

This leads to a final expression for  $V_r$  of

$$V_r = \frac{E\epsilon_r}{\rho c}$$

The shear strain in the specimen,  $\gamma_{rz}$ , is defined as

$$\gamma_{rz} = \frac{1}{2} \left( \frac{\partial u_r}{\partial z} + \frac{\partial u_z}{\partial r} \right)$$

where  $u_r$  is the displacement in the  $r$  direction and  $u_z$  is the displacement in the  $z$  direction. In the punch test,  $\partial u_r / \partial z = 0$ , and  $\partial u_z / \partial r \approx 0$ . Therefore, the shear strain can be approximated by

$$\gamma_{rz} = \frac{1}{2} \frac{\Delta u_z}{w} \quad (1)$$

The displacement at the end of the bar is defined as

$$u_z = \int (V_i + V_r) dt$$

Substituting in for  $V_i$  and  $V_r$ ,

$$\begin{aligned} u_z &= \int \left( -\frac{E\epsilon_i}{\rho c} + \frac{E\epsilon_r}{\rho c} \right) dt \\ &= \int \left( -\frac{c^2\epsilon_i}{c} + \frac{c^2\epsilon_r}{c} \right) dt \\ &= \int -c(\epsilon_i - \epsilon_r) dt \\ u_z &= -c \int (\epsilon_i - \epsilon_r) dt \end{aligned}$$

So now shear strain and strain rate become

$$\gamma_{rz} = -\frac{c}{2w} \int (\epsilon_i - \epsilon_r) dt \quad (2)$$

$$\dot{\gamma}_{rz} = -\frac{c}{2w} (\epsilon_i - \epsilon_r) \quad (3)$$

respectively.

The shear stress,  $\tau$ , is defined as the shear force divided the shear area where

$$\begin{aligned} A_{shear} &= \left( d_b + 2 \times \frac{w}{2} \right) \pi \times h \\ F_{shear} &= A_b E (\epsilon_i + \epsilon_r) = \frac{\pi d_b^2}{4} E (\epsilon_i + \epsilon_r) \end{aligned}$$

Combining the above equations yields

$$\tau_{rz} = \frac{d_b^2 E}{4h(d_b^2 + w)} (\epsilon_i + \epsilon_r) \quad (4)$$

Equations 2, 3, 4 are used to determine  $\tau_{rz}$ ,  $\gamma_{rz}$ , and  $\dot{\gamma}_{rz}$  for each test.

## 2.2 Explosives at AWEF

Experiments on actual explosives were conducted at the Advanced Warhead Experimentation Facility (AWEF) at Eglin AFB, Florida. This facility is in the Armament Branch of Wright Labs and has the facilities to conduct perform such tests safely with ease. Using a 12.7mm (0.50" caliber) powder gun, cylindrical steel projectiles 76.4mm in length were shot the specimens at velocities around 300 m/s. Pressure transducers with a known separation length at the end of the barrel produced voltage traces recorded on an oscilloscope, allowing a velocity to be calculated.

Both Tritonal and PBX-N109 specimens were made 36mm in diameter and 6.5mm thick. The specimens were placed in a removable insert which went into the die, see Figure 7. The inserts were made of 4340 steel hardened to Rockwell C of approximately 45. These inserts produced a clearance of 2.54mm between the projectile and inner diameter of the insert. Thin aluminum discs were placed behind the specimens to prevent loss of material due to spalling. Cover plates made of titanium 6% Al-4% V alloy were used in some shots to contain the explosive and determine the effect of a cover plate.

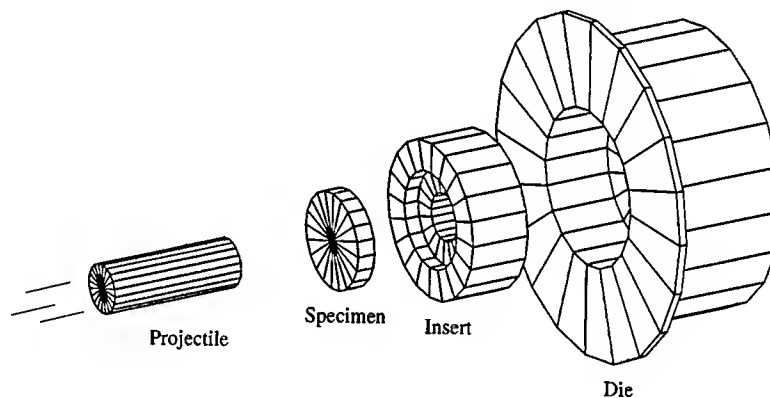


Figure 7: Die/insert configuration

The end of the powder gun and the specimen/die setup were all placed inside a steel-framed Lexan tank. The tank contained 2 feet of celotex behind the die to catch the projectile before impacting the back plate of the tank. A mirror was placed on one side of the die to allow two views of the event to be captured on each frame of film. A Cordin 330 high speed camera was placed on the opposite side of the mirror. See Figure 8. This allows for a side view and an angle view of the projectile impact in each frame. Framing rates were approximately 250,000/sec.

The use of high speed photography can play two important roles in investigating the processes mentioned above. First, the failure process can be examined. The amount of deformation and fracture, or a combination of the two, recorded on the film allows the determination of the failure mode under different loading cases.

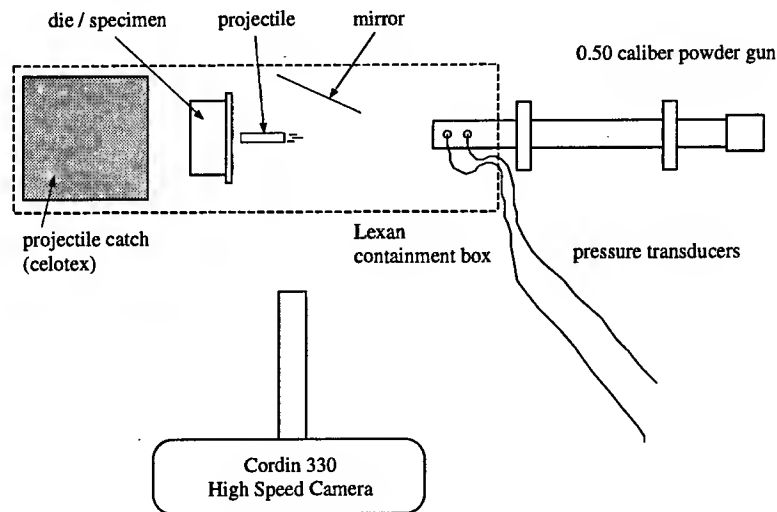


Figure 8: Experimental setup for test shots

Secondly, the initiation of the materials can be captured by eliminating any external light sources. The light generated by the reaction itself can be used to expose the film. This allows the amount of reaction, if any occurs, to be recorded. By examining the events captured on film, insight can be gained on how the mechanical and chemical processes within the reactive material interact together.

### 2.3 Hugoniot Analysis

As stated in the introduction, one method of initiation of explosives is by shock waves. Many experiments have been performed to characterize the shock response of explosives under different conditions. This experiment is examining the possibility of initiation through adiabatic shear localization or internal friction. To eliminate shock detonation as a possibility, the shock must be characterized for different impact velocities.

A one dimensional shock can be modeled as an instantaneous jump in a material from one state to another. This disturbance travels along at a speed  $U_s$ , see Figure 9. The shock wave changes the initial conditions,  $P_0$ ,  $\rho_0$ ,  $T_0$ , to the new conditions,  $P$ ,  $\rho$ , and  $T$  where  $P$  is the pressure,  $\rho$  is the density, and  $T$  is the temperature. The particle velocity also jumps from 0 to a certain value,  $U_p$ . By changing to a Lagrangian reference frame moving with the shock so the shock front appears stationary, the particles seem to approach the shock at a speed of  $U_s$  and leave with a velocity of  $(U_s - U_p)$ . This representation will be used in the impact problem described later.

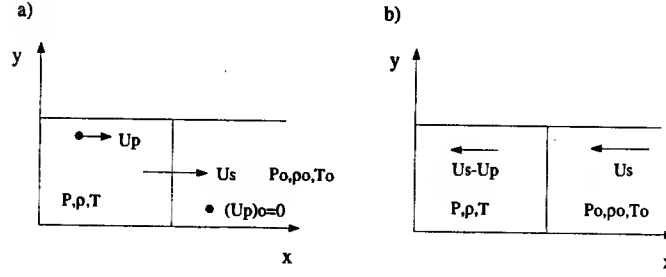


Figure 9: Schematic of a one dimensional shock from a) a stationary reference frame and b) a moving reference frame

The analysis begins with the conservation equations at the shock front.

Conservation of Mass:

$$\rho_0 U_s = \rho(U_s - U_p) \quad (5)$$

Conservation of Momentum:

$$(P - P_0) = \rho_0 U_s U_p \quad (6)$$

Conservation of Energy:

$$\begin{aligned} \Delta W &= (PA)(U_p dt) - (P_0 A)(U_0 dt) \\ \Delta E &= \frac{1}{2} [\rho A(U_s - U_p) dt] U_p^2 + EA \rho (U_s - U_p) dt \\ &\quad - \frac{1}{2} [\rho_0 A(U_s - U_0) dt] U_0^2 - E_0 A \rho_0 (U_s - U_0) dt \end{aligned}$$

where A is the cross-sectional area and  $U_0$  is the initial particle velocity. W is the work done on a particle as it passes through the shock front, and E is the energy in a material particle. Equating  $\Delta W$  to  $\Delta E$  and setting  $U_0=0$  yields

$$PU_p = \frac{1}{2} \rho (U_s - U_p) U_p^2 - E_0 \rho_0 U_s + E \rho (U_s - U_p)$$

Substituting the mass equation, equation (5), yields

$$PU_p = \frac{1}{2} \rho_0 U_s U_p^2 + \rho_0 U_s (E - E_0) \quad (7)$$

By using the mass and momentum equations, (5) and (6), the final form of the energy equation is

$$E - E_0 = \frac{1}{2} (P + P_0) (v_0 - v) \quad (8)$$

where  $v=1/\rho$ .

The four unknowns are  $U_s$ ,  $U_p$ , P, and E. With only 3 conservation equations, one more equation is needed for a complete set. This equation is called the equation of state (EOS). Though it is not the same



for a solid as an equation of state for a gas, it performs a similar role in the solution, so it has been given the same name<sup>1</sup>. This equation relates the shock and particle velocities through the following relation.

Equation of State (EOS):

$$U_s = C + SU_p \quad (9)$$

where C and S are material constants. These constants have been measured and tabulated in many references. By combining equations (6) and (9), a relation between the shock pressure and the particle speed can be obtained.

$$P = P_0 + \rho_0(C + SU_p)U_p \quad (10)$$

Equation (10) is very important in impact problems. The method of impedance matching can be used to find the shock pressure in the impact between two dissimilar materials [12]. During impact, the pressure at the interface must be equal. This gives rise to different shock and particle velocities by the equations above. The particles in the target travel at a velocity of  $U_{p2}$  while the projectile particles reduce speed by  $U_{p1}$  to a final value of  $V - U_{p1}$ , see Figure 10.

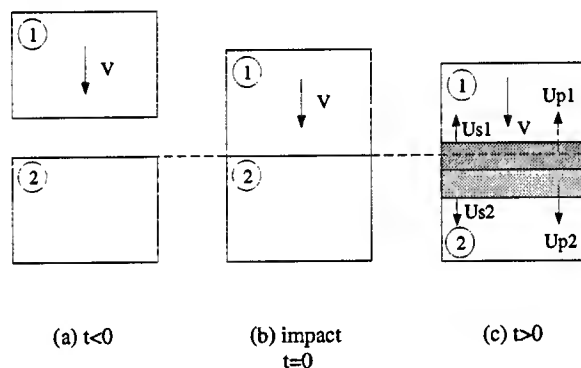


Figure 10: Schematic of impact problem

At the interface, the particle speeds must be the same, so

$$V - U_{p1} = U_{p2}$$

and

$$U_{p2} + U_{p1} = V$$

Conservation of momentum and the equations of state for the projectile and target can be used to find the final values. The conservation of momentum for the projectile and target are

$$P_1 = \rho_{01}U_{s1}U_{p1}$$

<sup>1</sup>The more conventional EOS for gases is  $E = Pv/(\gamma - 1) - \lambda q$  where  $\lambda$  and  $q$  are reaction parameters. This equation has been used for solids with a constant  $\gamma = 3$  [9] which allows solution for detonation wave speeds and reaction zone structures, but is not suitable for the impact problem.

$$P_2 = \rho_{02} U_{s2} U_{p2}$$

respectively.  $P_0$  is assumed to be 0. This is a good assumption as pressures can routinely reach above 100 atmospheres in these analyses. The EOS for the two materials are

$$U_{s1} = C_1 + S_1 U_{p1}$$

$$U_{s2} = C_2 + S_2 U_{p2}$$

Combining the equations yields

$$P_1 = \rho_{01}(C_1 + S_1 U_{p1})U_{p1} = \rho_{01}C_1 U_{p1} + \rho_{01}S_1 U_{p1}^2$$

$$P_2 = \rho_{02}(C_2 + S_2 U_{p2})U_{p2}$$

Using the substitution  $U_{p1} = V - U_{p2}$  and setting  $P_1 = P_2$ , the following quadratic equation for  $U_{p2}$  can be derived.

$$U_{p2}^2(\rho_{02}S_2 - \rho_{01}S_1) + U_{p2}(\rho_{02}C_2 + \rho_{01}C_1 + 2\rho_{01}S_1V) - \rho_{01}(C_1V + S_1V^2) = 0$$

The roots of the equation are

$$U_{p2} = \frac{-(\rho_{02}C_2 + \rho_{01}C_1 + 2\rho_{01}S_1V) \pm \sqrt{\Delta}}{2(\rho_{02}S_2 - \rho_{01}S_1)} \quad (11)$$

where

$$\Delta = (\rho_{02}C_2 + \rho_{01}C_1 + 2\rho_{01}S_1V)^2 - 4(\rho_{02}S_2 - \rho_{01}S_1)(-\rho_{01})(C_1V + S_1V^2)$$

Two solutions for  $U_{p2}$  will be found due to the quadratic nature of the equation. The determination for which solution to use is governed by the fact that  $U_{p2}$  must be less than the original projectile velocity,  $V$ . With  $U_{p2}$  known, all other quantities, including the shock pressure  $P_2$ , can then be calculated with the relations already discussed.

A graphical solution of the same problem is also possible. By plotting pressure versus particle velocity for the projectile and target, the solution can be read directly. Plot the target curve normally, and then plot the projectile curve with the origin at  $V$ , and then invert the curve (change  $U_p$  to  $-U_p$ ). Where the two curves cross is the solution to the impact problem. Figure 11 shows the  $P$ - $U_p$  plane for a steel on TNT impact. TNT was used because its behavior is close to that of Tritonal, and the data was readily available. The impact velocity of 1650 m/s was used to give a shock pressure  $P$  of 10.7 GPa. Particle speeds in the TNT are 325 m/s with a shock speed of 4200 m/s. The same parameters for steel are 1325 m/s and 4660 m/s, respectively.

The impact speed of 1650 m/s was chosen above to give the final shock pressure of 10.7 GPa. This pressure is the initiation pressure for TNT [7] for this geometry. This speed is much higher than the 300

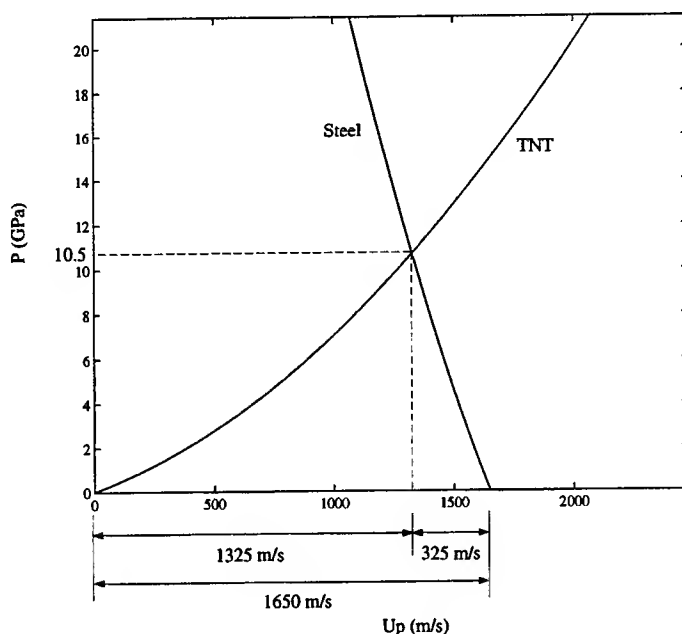


Figure 11: Hugoniot analysis of steel on TNT impact

m/s velocity used in the experiment. Because the velocities are so much lower, shock initiation can be ruled out as a possible mechanism for reaction during the tests.

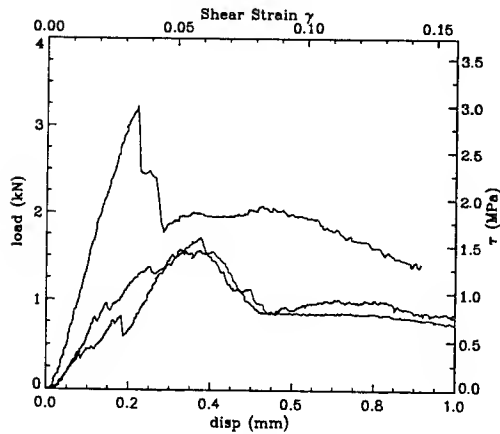
### 3 Results & Discussion

#### 3.1 Punch Tests on Explosive Simulants

The load-displacement curves for the simulants under quasi-static loading rates are shown in Figure 12. For both the Filler-E and the cure cast simulant, there was one test of the three that showed a slightly different behavior. This shows the variety of material properties within these materials. Repetitive behavior may be hard to obtain due to the numerous possibilities for imperfections in the materials. Post mortem examination revealed that the specimens did repeat failure behavior qualitatively. The Filler-E was very brittle. A central plug formed, with radial cracks extending outward to the outer diameter. The cure cast PBX simulant formed a plug every time, and did not have any radial fractures. There were no fragments, leading to the conclusion that the material tore. Average stress- strain curves for the two materials are shown in the same graphs.

At low velocities on the mechanical press, there is not much change in the behavior of the materials. Both Filler-E and the cure cast PBX simulant showed the same failure pattern as in the quasi-static tests. The maximum loads in the materials were higher, showing strain rate hardening, which is very large in the case of the cure cast PBX simulant. Failure strains for both the cure cast PBX simulant and Filler-E reduced by

a)



b)

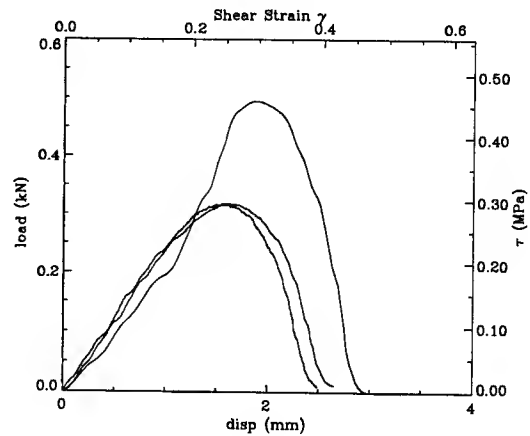
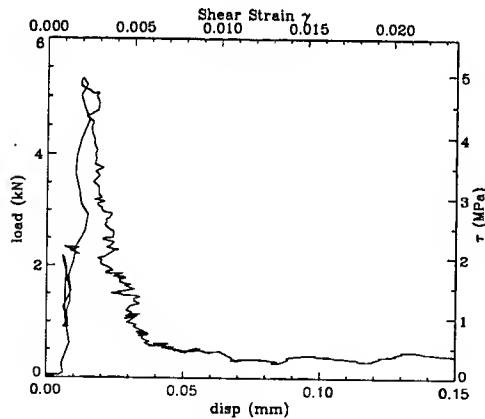


Figure 12: Quasi-static load-displacement and average stress-average strain curves for (a) Filler-E and (b) cure cast PBX simulant

a factor of 10 from the quasi-static tests. In the case of the Filler-E, this is most likely due to fracture as the compressive wave reflects of the back face and becomes tensile. Due to the brittle nature of Filler-E, the specimen cracks very quickly under the dynamic loading. The PBX simulant seems to be exhibiting a large strain rate dependence in both load capacity and failure strain. Load-displacement and average stress-strain curves from the mechanical press are in Figure 13.

a)



b)

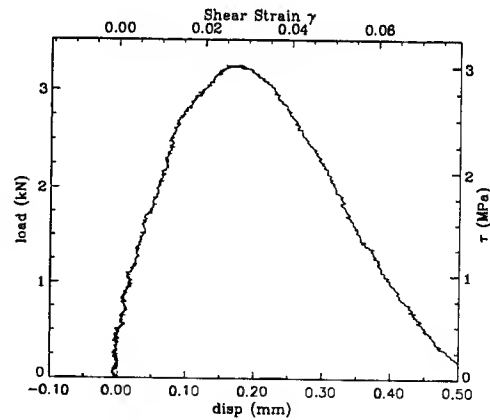


Figure 13: Low velocity load-displacement and average stress-average strain curves for (a) Filler-E and (b) cure cast PBX simulant

Results for a Hopkinson bar tests with no specimen allows for a verification of the method, as well as giving insight into what results can be expected during an actual test. Figure 14 shows the raw data obtained from the digital oscilloscope. The top trace shows both strain gage histories, and the second shows the average.

Averaging the two eliminates any bending that may occur. Bending should be kept to a minimum, which the first trace shows occurring (the two strain gages curves lie almost directly on top of each other). If large amounts of bending are present, the impact of the bar with the specimen will not be perpendicular and may lead to erroneous results. The two strain gages have very similar voltage traces, verifying the absence of a large bending stress. The impact speed in this test was approximately 11.5 m/s.

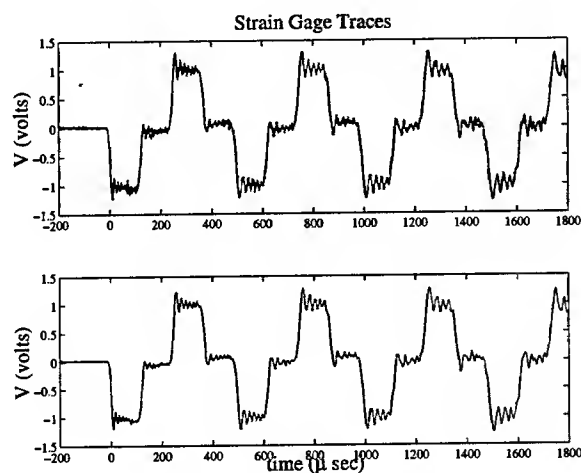


Figure 14: Voltage traces for Hopkinson bar

By placing the initial compressive and tension pulses over each other, another aspect of this experimental method is revealed. With no specimen present, the two pulses should have the same shape. However, wave dispersion has caused the pulses to change shape [11]. Wave dispersion describes the phenomenon that in uniaxial stress pulses, waves of different frequencies travel at different velocities. As a Fourier analysis will show, a square pulse is made up of many different waves of various frequencies and amplitudes. These different waves start to diverge immediately, and thus the shape of the wave changes as it travels. The compressive pulse (plotted as tension) and the reflected tension pulse can be seen in Figure 15. The magnitudes of the strain pulses,  $1100\mu\epsilon$ , verifies that the data follows the elastodynamic relation  $V = E\epsilon/\rho c$  derived in section 2.1.1.

Due to the changing of the shape, noise is generated in the data calculated from the pulses. Figure 16 shows load, stress, strain rate, and displacement at the end of the bar during the pass of the initial pulse. The load and stress should be zero for all time. The strain rate is the average strain rate that a specimen would see as determined by the displacement seen in the final graph. The magnitude of the displacement is as expected from knowing the length of the pulse, the elastic wave speed, and the particle velocity in the wave.

The first noticeable error is in the load graph. The load is obviously not zero due to the change in shape of the pulses caused by dispersion. Similarly, even though the end of the bar must physically be a stress

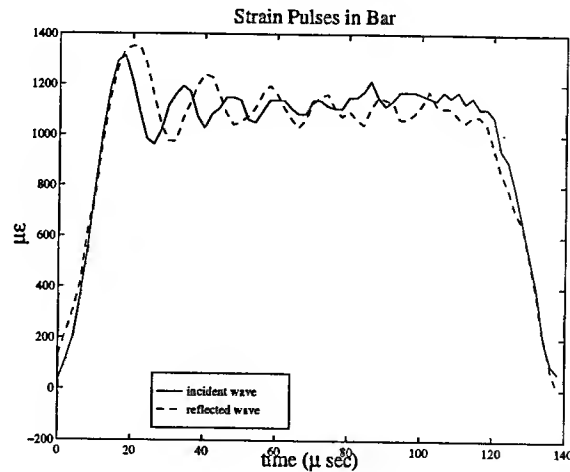


Figure 15: Comparison of initial and reflected strain pulses

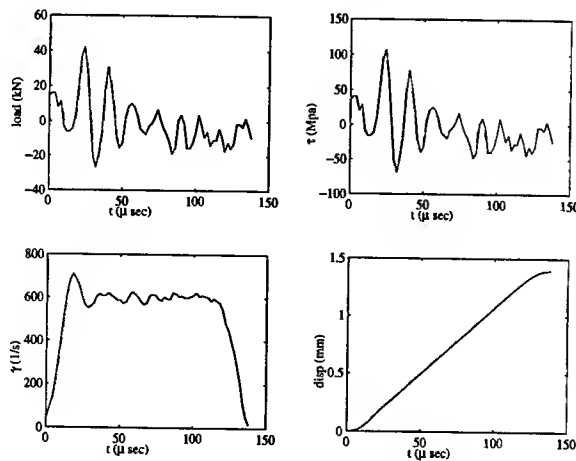


Figure 16: Data calculated from strain pulses

free surface, the stress is reported to be non-zero. The distressing part about the noise is the magnitude of the load. The load repeatedly reaches values of  $\pm 20$  kN. While such an uncertainty may be acceptable when testing metals, this is well above the expected load required to fail either of the inert simulants and is unacceptable. The maximum loads reported so far for failure are about 6 kN.

Results of the Hopkinson bar testing are shown below in Figure 17. Post mortem examination showed no difference in failure mode from the tests conducted on the mechanical press. The impact velocities for the Filler-E and PBX simulant are 11.04 and 12.35 m/s, respectively. The results do not reveal much about the behavior of the simulants at these punch speeds. The error from dispersion overwhelms any data that may be present in the voltage traces. These simulants are so much weaker than any steel used that the strength of the simulants, or the explosives they model, may be neglected. Similar tests on 1018 steel give yield strengths of 300 MPa. This technique is better used for metals which have a much higher yield strength

and elastic modulus.

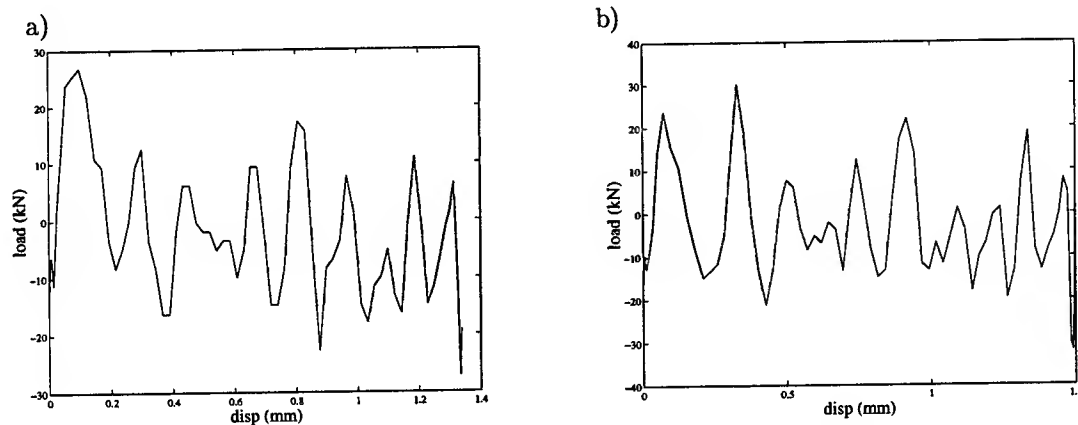


Figure 17: High velocity load-displacement curves for a) Filler E and b) cure cast simulant

### 3.2 Impact of Tritonal and PBX N109

With the fracture characterized for lower velocities, the experiments at the AWEF reveal the failure behavior at very high loading rates and their effect of the initiation on the materials. As can be seen in the pictures taken with the Cordin camera (Figures 18 and 19), both the materials fracture quickly and are ejected from the insert. This ejected material is not reacting at any appreciable rate. Both the materials behave similarly, but the Tritonal seems to have more material ejected. Tritonal is a brittle material that fractures quite easily. With the initial impact, the Tritonal forms fragments which are then ejected with further projectile motion. The PBX-N109, on the other hand, is a weak but more ductile material. It offers little resistance to the projectile motion, but is ejected in larger pieces from the insert.

None of the test shots caused detonation of the specimens. The pressure generated in the shock wave by the impact is 1.23 GPa for the Tritonal, and even less for the PBX-N109 as determined by the Hugoniot method [12]. From Pop plots for TNT [7], these pressures will not cause shock detonation for a specimen 0.25" thick. Limits of the gun prevented higher impact speeds.

In other tests, a cover plate of titanium 6% Al-4% V alloy was used to determine if containment and possible heating from the plate would cause detonation. Again, there was no reaction.

To get a detonation, the reactive material must be contained and compressed [3]. In this experiment, it is too easy for the fractured material to escape; there is negligible internal friction and negligible hydrostatic pressure in the shear zone. High pressure combined with shear friction could easily lead to ignition. Boyle et al. [3] performed tests that combined pressure *and* shear showing that TNT would ignite for velocities around 80 m/s with a pressure of 0.5 GPa. The method of initiation in Boyle's tests is probably internal friction as the tests with steel on explosive and explosive on explosive boundaries made no difference in

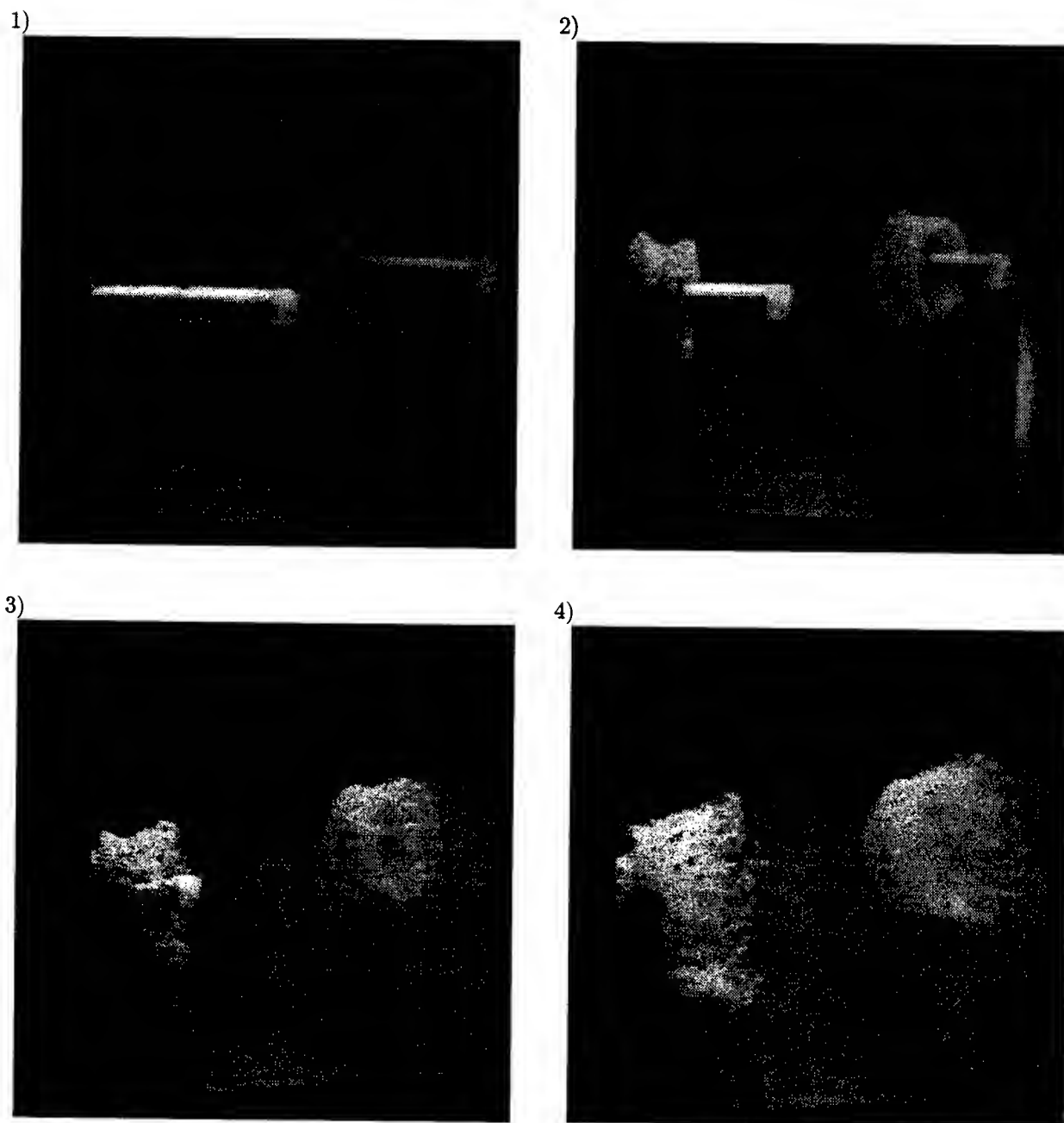


Figure 18: High speed photography of PBX N109 impact, approximate times after impact are: 1)  $36\mu s$ , 2)  $116\mu s$ , 3)  $180\mu s$ , and 4)  $272\mu s$



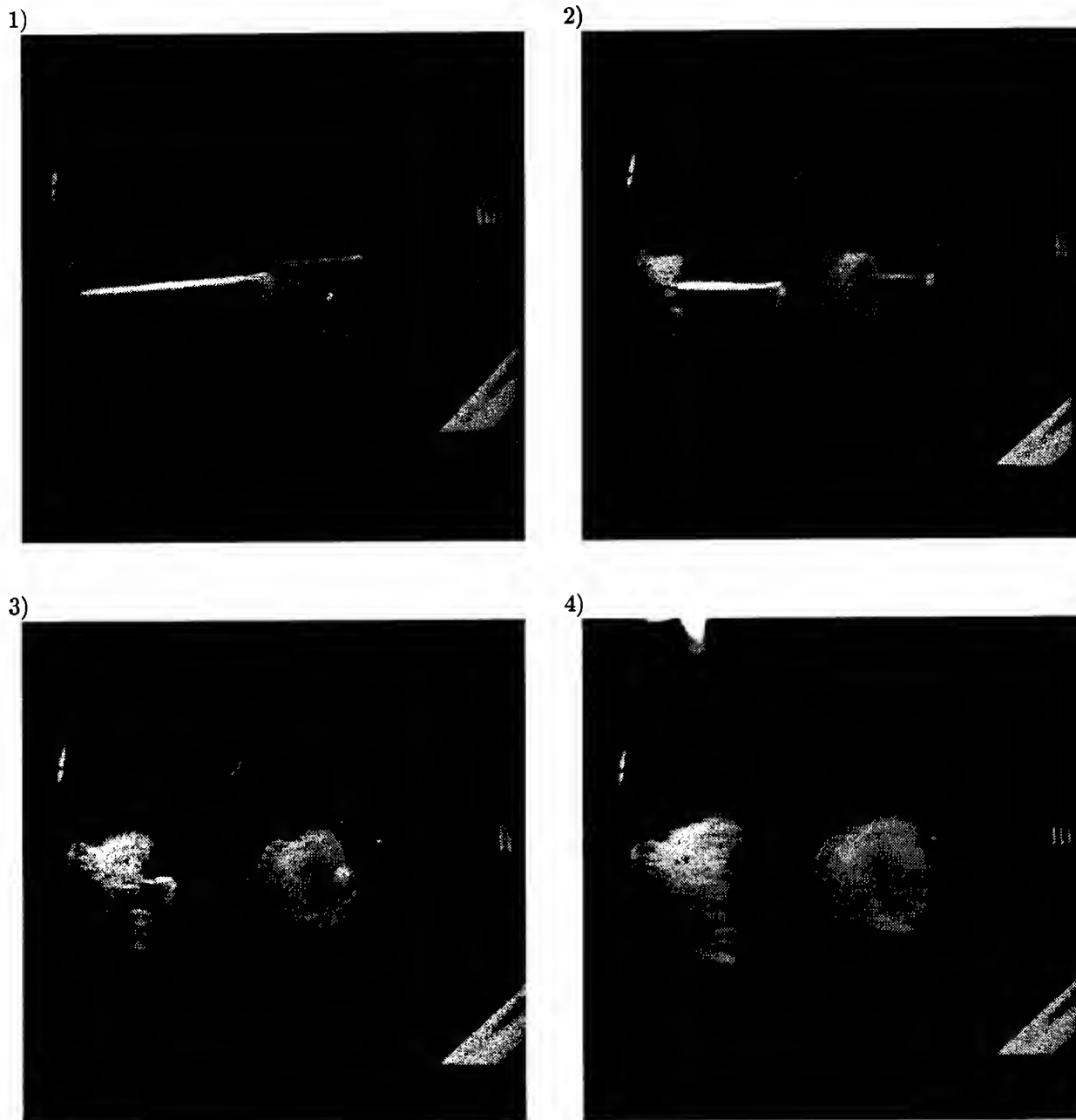


Figure 19: High speed photography of Tritonal impact, approximate times after impact are: 1)  $28\mu\text{s}$ , 2)  $112\mu\text{s}$ , 3)  $184\mu\text{s}$ , and 4)  $252\mu\text{s}$

the outcome. The main difference between these experiments and Boyle's is that in Boyle's tests there is an applied hydrostatic pressure; clearly the existence of a hydrostatic pressure is critical to the dominant ignition mechanism in that case.

During the experiments conducted at the AWEF, the technique of turning off external lights to capture light given off by the reaction was not used. Since very little, if any, reaction occurred, it is expected that no light was emitted.

Chou [5] has run numerical models on similar experiments, one of which included an impact of TNT by a steel projectile at 200 m/s. In his calculations, he shows that a shear band forms at the corner of the projectile. Temperatures reach 500°F. These temperatures are not high enough within the band to cause ignition. At impact velocities of 1000 m/s, pressures and temperatures become high enough to cause detonation for his model. What Chou does not take into account is the fracture of the material. As seen in the photographs, the Tritonal fractures very quickly. Even with cover plates, the materials is no longer a continuous solid, but a granular material and should be modeled appropriately. Streak camera photographs have shown that detonation waves in pressed or cast explosives are rough, indicating that the flow is irregular in its fine detail. Initiation thus depends strongly on the type, number, and distribution of inhomogeneities [4]. By fracturing these materials, the initiation behavior may change dramatically.

## 4 Conclusions

For both Filler-E and the cure cast PBX simulant, the ultimate strengths of these materials are far below anything used for a bomb casing, usually 4340 steel. When modeling the mechanical strength of the weapon as a whole, the explosive can be neglected as it will not affect the outer casing's behavior due to its negligible relative strength.

From the behavior of the Filler-E, it can be inferred that that upon impact, Tritonal will fracture very quickly. In a deep earth penetrator, the explosive may even become a granular material before ignition occurs. This may hamper or improve ignition of the explosive depending on loading conditions. Penetration of the weapon by a foreign object could have disastrous effects. The fracturing and then compression of Tritonal could lead to ignition. Internal friction is greatly increased after the fracture occurs if a large hydrostatic pressure is present. All the free surfaces are then allowed to rub against each other, greatly increasing the risk of premature initiation.

The cure cast PBX, which does not fracture at lower loading rates, also fails into many small particles upon impact. The cure cast simulant showed a large amount of strain rate hardening and decrease in failure strain between the quasi-static and low velocity tests, and this trend seems to continue to the very high

velocities for PBX N109. Though load carrying capacity at this loading rate is not known, the failure mode is much closer to that of the Tritonal. Again, the effect on the initiation from any failure is unknown, but internal friction with a large hydrostatic pressure will be the primary cause of initiation at these loading rates.

## Nomenclature

$r$	radial direction
$z$	direction along punch axis
$u_r$	displacement in $r$ direction
$u_z$	displacement in $z$ direction
$v$	specific volume in Hugoniot analysis
$w$	shear width (clearance)
$d_b$	bar diameter
$A_b$	bar area
$h$	specimen thickness
$c$	sound velocity in punch
$E$	modulus of elasticity in punch, energy in Hugoniot analysis
$\epsilon$	axial strain in punch
$\tau$	shear stress in specimen
$\gamma$	shear strain in specimen
$\dot{\gamma}$	shear strain rate in specimen
$V_i, V_r$	incident and particle velocity in Kolsky bar
$U_s$	shock velocity
$U_p$	particle velocity in projectile or target for impact problems
$P$	pressure
$T$	temperature
$\rho$	density
$W$	work

## References

- [1] ASM International. *ASM Metals Handbook*, 10th edition, 1990.
- [2] F.P. Bowden and Y.D. Yoffe. *Initiation and Growth of Explosives in Liquids and Solids*. Cambridge University Press, New York, 1952.
- [3] V. Boyle, R.B. Frey, and O. Blake. Combined pressure shear ignition of explosives. *9th International Symposium on Detonation*, 1989.
- [4] A.W. Campbell, W.C. Davis, J.B. Ramsay, and J.R. Travis. Shock initiation of solid explosives. *The Physics of Fluids*, 4(4):511-521, April 1961.
- [5] P.C. Chou. Explosive response to unplanned stimuli. Final technical report, Dyna East Corporation, Philadelphia, PA, September 1991. submitted to Armament Directorate, Eglin AFB, FL.
- [6] W.C. Davis. High explosives: The interaction of chemistry and mechanics. *Los Alamos Science*, 2:48-75, 1981.
- [7] B.M. Dobratz and P.C. Crawford. *LLNL Explosives Handbook - Properties of Chemical Explosives and Explosive Simulants*. National Technical Information Service, Springfield, VA, January 1985. Document # DE91006884.
- [8] A.R. Dowling, J. Harding, and J.D. Campbell. The dynamic punching of metals. *J. Inst. Metals*, 98:251-224, 1970.
- [9] W. Fickett and W.C. Davis. *Detonation*. University of California Press, Berkeley, CA, 1979.
- [10] J.E. Field, G.M. Swallowe, and S.N. Heavens. Ignition mechanisms of explosives during mechanical deformation. *Proceedings. Royal Society of London*, A 382:231-244, 1982.
- [11] K. F. Graff. *Wave Motion in Elastic Solids*. Dover Publications, Inc., New York, 1975.
- [12] Marc A. Meyers. *Dynamic Behavior of Materials*. John Wiley & Sons, Inc., New York, 1994.
- [13] A.K. Zurek. The study of adiabatic shear band instability in a pearlitic 4340 steel using a dynamic punch test. *Metallurgical and Material Transactions A*, 25A:2483-2489, 1994.